Development of Methods for Protein Delivery and the Directed Evolution of Recombinases

Development of Methods for Protein Delivery and the Directed Evolution of Recombinases

A dissertation presented

by

David Brandon Thompson

to

The Division of Medical Sciences

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

April 2014

Dissertation Advisor: David R. Liu David Brandon Thompson

Development of Methods for Protein Delivery and the Directed Evolution of Recombinases

Abstract

As a class, protein-based therapeutics offer tremendous advantages over traditional small molecule drugs. Due to their sizes and folding energies, proteins are ideal for catalyzing chemical reactions, and can bind tightly and selectively to extended target surfaces. However, due to their large size, virtually all proteins are unable to spontaneously enter cells, and as a result protein therapeutics are restricted to extracellular targets. We developed a platform for delivery of proteins to intracellular target sites by engineering the surface chemistry of a model protein, green fluorescent protein (GFP). We found that 'supercharged' cationic GFP variants (scGFPs) bind to anionic cell surface molecules and initiate endocytosis, resulting in the efficient delivery of translationally fused cargo to intracellular targets. We discovered that scGFPs, and cationic delivery reagents in general, alter endosomal trafficking in a manner proportional to both their charge and their delivery efficiency, suggesting that avoidance of endosomal maturation is a key step in the endosomal escape of delivered protein cargos. We also developed a method for encapsulation of recombinant proteins by cationic lipid delivery reagents using negatively supercharged GFP.

Genetic modification technologies have matured rapidly following the discovery of protein classes with programmable DNA-binding specificities. While site-directed genetic knockout technologies are highly effective, targeted integration and repair remain comparatively inefficient. Site-specific recombinases directly catalyze strand exchange and ligation between DNA molecules, offering an approach to efficient genomic integration. However, most site-

specific recombinases are not easily reprogrammable. To address this problem, we developed a genetic selection technique based on the Phage-Assisted Continuous Evolution (PACE) system, to enable the rapid evolution of recombinase proteins towards targets of interest. Using Cre recombinase as a model, the PACE system was optimized, validated, and used to evolve Cre variants with higher activity on their native loxP target site, as well as altered specificity towards a human genomic sequence within the hROSA26 locus.

Finally, we developed a method for enhancing the specificity of RNA-guided nucleases by restricting activity to sites of obligate dimeric nuclease assembly. We engineered a *Fok*I nuclease fusion to a catalytically inactivated Cas9 protein that mediates efficient modification with significantly reduced off-target activity.

Table of Contents

Acknowledgements

My thesis work would not have been possible without the support and guidance of my advisor, Dr. David R. Liu. David has provided me with a research environment that encourages independence and self-starting pursuit of science. The extreme degree of freedom afforded in this environment has had an undeniable impact on my development as a scientist. I have been given the latitude to participate in a wide range of distinct research projects over the course of my thesis. Thanks to David's guidance and the environment fostered in his lab I have been able to pursue each area effectively and successfully. Thanks to David, I have also had an exceptional amount of hands-on exposure to the process of developing research manuscripts and grant applications, skills I expect to be invaluable to me in my future career.

I would like to thank the members of my dissertation advisory committee, Dr. Stephen J. Elledge, Dr. Susan Lindquist, and the chair of my committee, Dr. J. Keith Joung, for their continual scientific and career guidance throughout my time in graduate school. I would also like to thank Dr. Alan Saghatelian, Dr. Timothy K. Lu, and Dr. Jagesh V. Shah for their time and consideration of this thesis and for serving as my dissertation defense examiners.

I am thankful for the opportunities I have had over the course of my graduate career to work with many skilled scientific collaborators. Dr. Brian McNaughton initiated the area of supercharged protein delivery in the Liu lab, enabling a major portion of my own research in the process. Dr. James Cronican and I worked closely on the initial development of supercharged protein delivery techniques, and James was instrumental in success of this collaborative work. Dr. Kevin Beier, a member of Dr. Constance Cepko's laboratory at Harvard Medical School, worked tirelessly in demonstrating the efficacy of our supercharged protein delivery technology

in vivo using a mouse retinal model. Brent Dorr assisted with computational modelling of proteins during my study of the structural and functional properties of supercharged proteins. Dr. Roberto Villaseñor, a member of Dr. Marino Zerial's laboratory at the Max Planck Institute of Molecular Cell Biology and Genetics, assisted in the acquisition and analysis of high-throughput confocal microscopy data during the study of supercharged protein uptake and intracellular trafficking. Dr. John Zuris and I collaborated on the initial validation of a cationic lipid-based recombinant protein delivery method I conceived of, and John has been central to the development of this platform for broader applications. Dr. John Guilinger and I collaborated closely during the final months of my thesis work on the development of the *Fok*I-dCas9 genome-editing toolset, sharing equally in the experimental design, execution, and analysis of our research. I would also like to thank Dr. Jacob Carlson and Ahmed Badran for their knowledge, advice, and materials graciously provided during my work developing the Phage-Assisted Continuous Evolution system for site-specific recombinases.

I thank my mother and father, Doris and David, for their unending love and support in everything I do. They have given me a sense of curiosity about the world as it exists, and a sense of imagination and creativity for what it might become. Together, they have taught me to work diligently and honestly, to always aim high, and to never lose sight of the big picture. I also thank my sisters, Desirée and Danika, for their love and support. Finally I would like to thank my graduate school friends and the members of the Liu lab for their support, conversation, and laughs.

Chapter 1:

Development of an in vivo and in vitro Protein Delivery Platform using Supercharged Green Fluorescent Protein

Abstract

The inability of proteins to potently penetrate mammalian cells limits their usefulness as tools and therapeutics. When fused to superpositively charged GFP, proteins rapidly (within minutes) entered five different types of mammalian cells with up to ~100-fold greater potency than corresponding fusions with known protein transduction domains (PTDs) including Tat, oligoarginine, and penetratin. Ubiquitin-fused supercharged GFP when incubated with human cells was partially deubiquitinated, suggesting that proteins delivered with supercharged GFP can access the cytosol. Likewise, supercharged GFP delivered functional, non-endosomal recombinase enzyme with greater efficiencies than PTDs *in vitro*, and also delivered functional recombinase enzyme to the retinas of mice when injected *in vivo*.

Introduction

Proteins have demonstrated great value as research tools and as human therapeutics. Due to the inability of virtually all proteins to spontaneously enter cells, however, exogenous proteins are predominantly restricted to interaction with extracellular targets and targets accessible through the endocytic pathway. Over the past decade, a variety of reagents for the delivery of proteins into mammalian cells have been developed including lipid-linked compounds,[1] nanoparticles,[2] and fusions to receptor ligands.[3, 4] Perhaps the most commonly used method for protein delivery is genetic fusion to PTDs including the HIV-1 transactivator of transcription (Tat) peptide, oligoarginine, and the *Drosophila* Antennapedia-derived penetratin peptide.[5, 6] Despite these advances, intracellular targets remain difficult to perturb using exogenous proteins; even modest success can require high concentrations of exogenous protein due to the modest potency of most current methods. Challenges for protein delivery are significantly increased *in vivo*, where cells in the context of a live animal have proven especially difficult targets for

functional protein delivery.[7, 8] The development of more potent protein transduction platforms would therefore significantly increase the scope of potential applications for protein reagents and therapeutics.

We recently described "supercharged" GFP variants that have been extensively mutated at their surface-exposed residues, resulting in extremely high theoretical net charge magnitudes ranging from –30 to +48 (Figure 1.1).[9] We discovered that superpositively charged GFP variants can enter a variety of mammalian cells by binding to anionic cell-surface proteoglycans and undergoing endocytosis in an energy-dependent and clathrin-independent fashion.[10] Further, we observed that superpositive GFPs are able to form stable non-covalent complexes with nucleic acids and that +36 GFP can deliver siRNA and plasmid DNA into a variety of mammalian cell lines without apparent cytotoxicity.

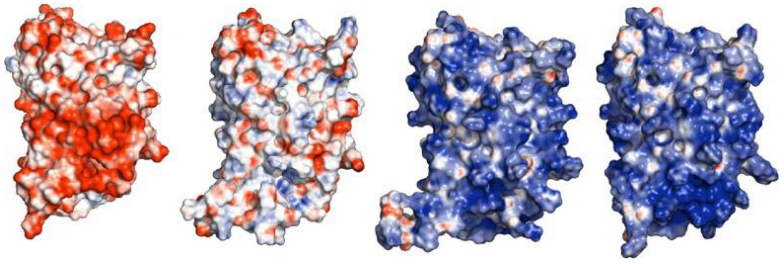

Figure 1.1 Electrostatic surface potentials of supercharged GFPs. -30 GFP, stGFP (the starting, non-supercharged form of GFP, +36 GFP, and +48 GFP colored from -25kT/e (red) to +25kT/e (blue).

Results

We hypothesized that +36 GFP may also serve as a potent and general platform for the delivery of proteins into mammalian cells. We began by generating a variety of fusion proteins with +36 GFP, including the Tat peptide tag, ubiquitin, mCherry, and Cre recombinase. We observed that each fusion retained the fluorescence excitation and emission spectra of the unfused GFP (Figure 1.2), enabling measurement of their cell penetration by detection of cellular

fluorescence. The purified fusion proteins maintain the ability to rapidly (within 15 min) and potently (at low nM concentrations) penetrate mammalian cells (Figure 1.3) with low toxicity at doses in the low micromolar range (Figure 1.4).

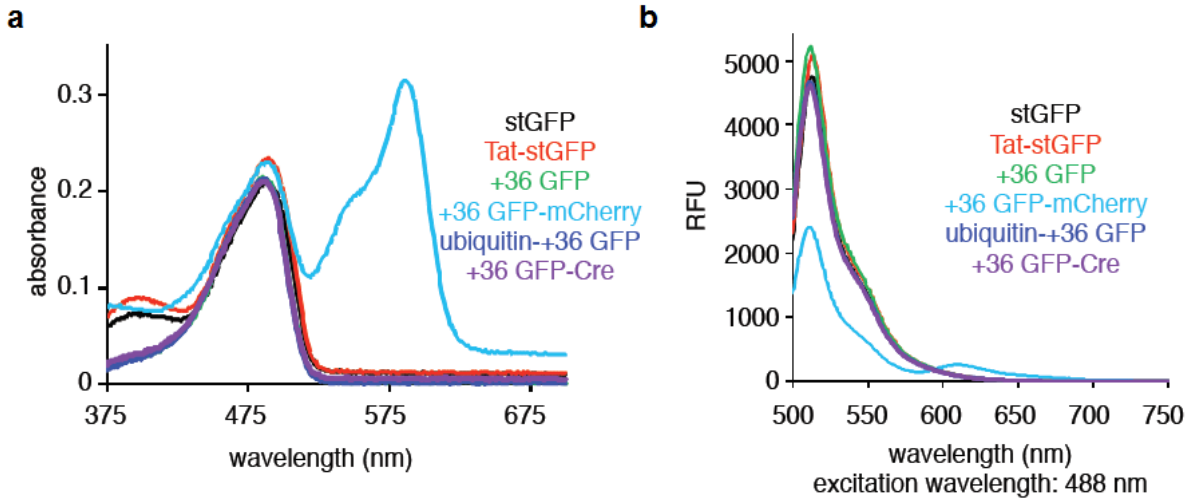

Figure 1.2 Characterization of +36 GFP fusion proteins. (**a**) Absorbance and (**b**) emission spectra of GFP fusion proteins shows equal fluorescence except for +36 GFP mCherry which shows reduced fluorescence at 515 nm and also an extra emission peak at 615 nm, presumably due to FRET with the attached mCherry fluorophore.

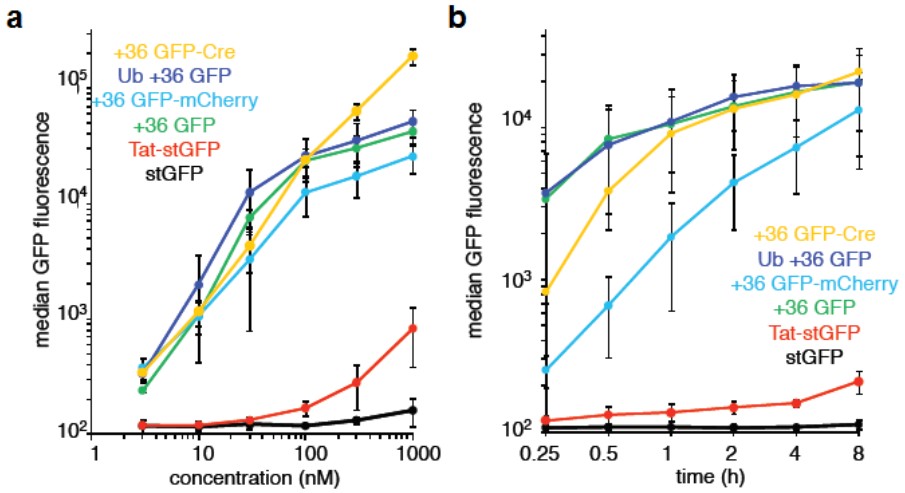

Figure 1.3 +36 GFP protein fusions penetrate cells rapidly and potently. (a) Flow cytometry of HeLa cells incubated at the concentrations shown in the presence of +36 GFP fusions for 4 hours at 37° C. Cells were washed three times with 20 U/mL heparin in PBS to remove membrane bound protein prior to analysis. Untreated cells resulted in median GFP fluorescence values of 107 ±5. Error bars represent the standard deviation of three independent biological replicates. **(b)** Flow cytometry of HeLa cells incubated in the presence of 100 nM of each +36 GFP fusion at 37° C for the specified time. Untreated cells resulted in median GFP fluorescence values of 100 ± 6. Error bars represent the standard deviation of three independent biological replicates.

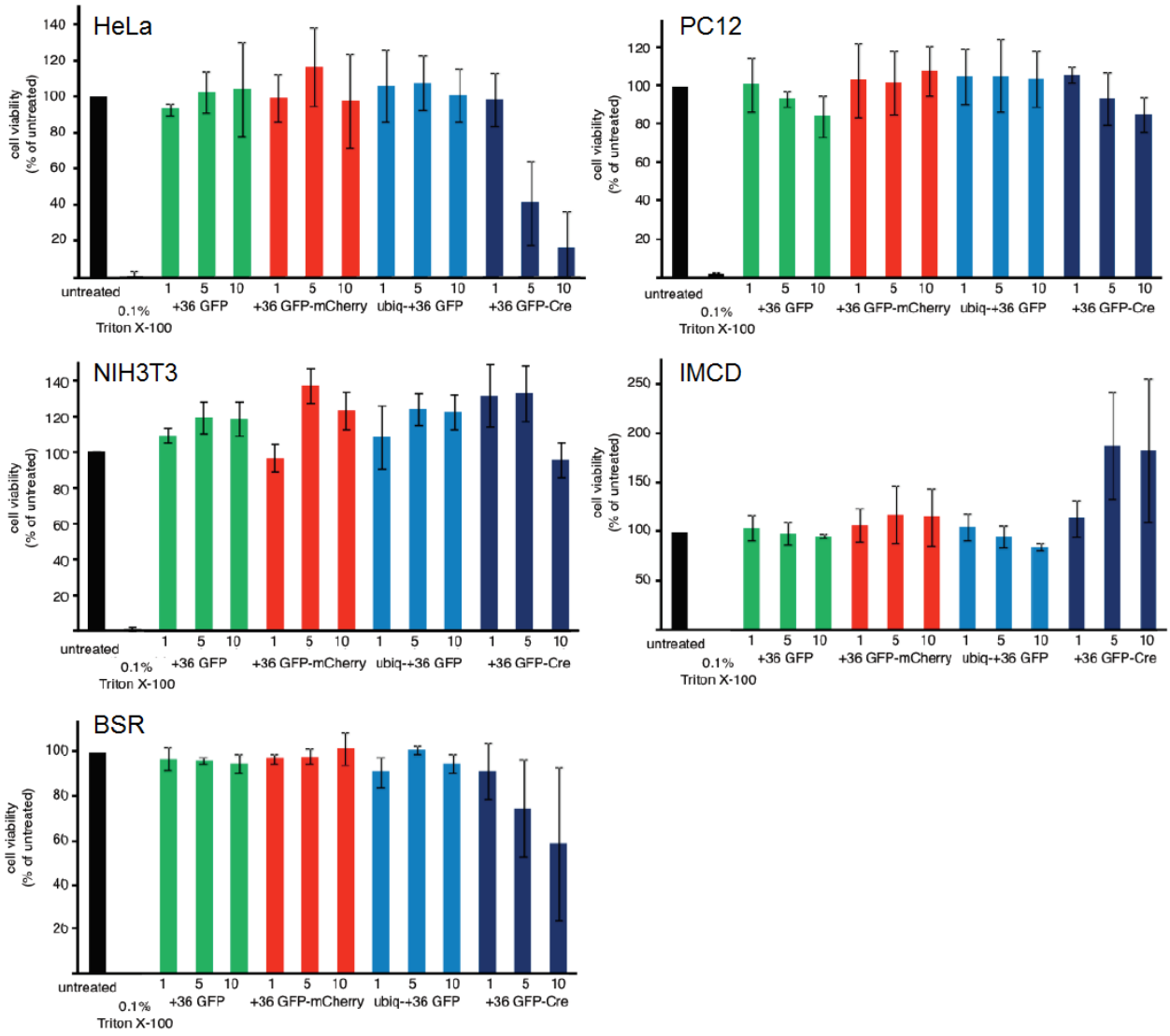

Figure 1.4 +36 GFP and +36 GFP fusions are not toxic at concentrations effective for protein delivery. At concentrations ≥ ~10 to 100 times the effective concentration for protein delivery in this work, +36 GFP-Cre (but not other +36 GFP fusion proteins or +36 GFP itself) reduced the viability of some cell lines and possibly stimulated IMCD cells. Values and error bars represent the average of and standard deviation, respectively, of three independent biological replicates.

Next we directly compared the ability of +36-GFP, Tat, Arg_{10}, and penetratin to deliver fused mCherry,[11] a red fluorescent protein variant. We generated +36 GFP-mCherry, Tat-mCherry, Arg_{10}-mCherry, and penetratin-mCherry with identical linkers and fusion orientations and found that their fluorescence properties were remarkably similar, enabling direct comparison of cell penetration between fusion proteins (Figure 1.5). We incubated these four fusion proteins

with HeLa cells, baby hamster kidney cells (BSR cells, a clone of BHK-21 cell line), NIH 3T3 cells, inner medullary collecting duct (IMCD) cells, and rat pheochromocytoma PC12 cells in serum-free media at various concentrations for 4 h at 37 °C.

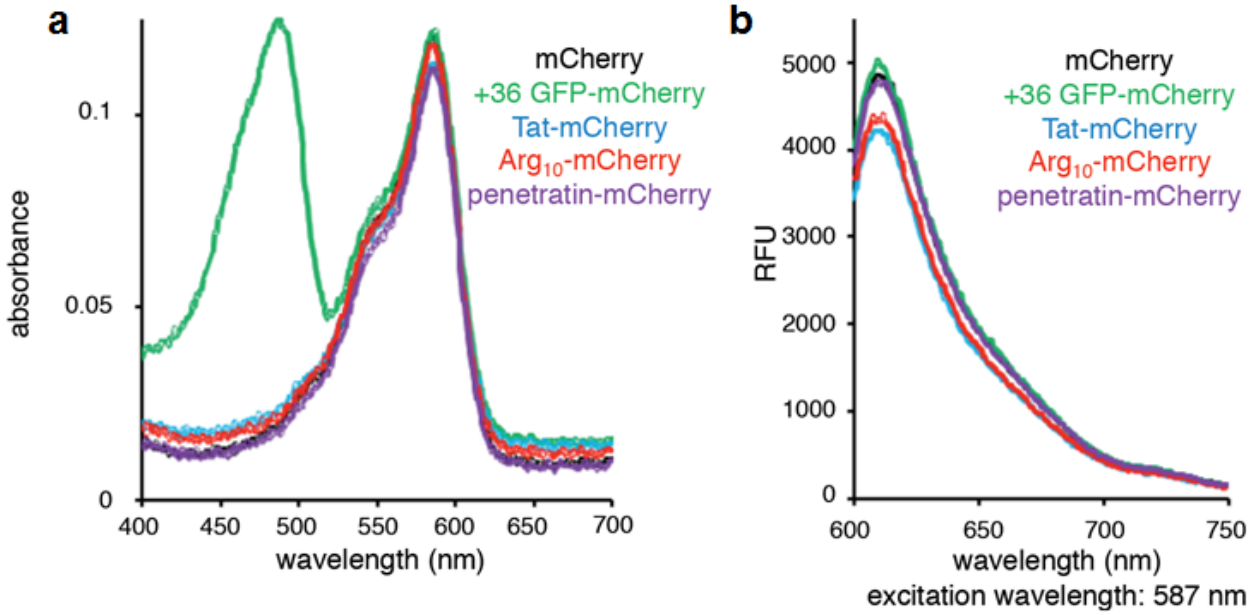

Figure 1.5 Characterization of mCherry fusion proteins. Analysis of mCherry fusions to Tat, Arg10, and penetratin by **(a)** absorbance from 400 nm to 700 nm, and (**b**) emission spectra at 587 nm excitation.

After incubation, cells were washed under conditions confirmed to remove surface-bound protein (Figure 1.6), trypsinized, and assayed for internalized mCherry by flow cytometry (Figure 1.7). For all five cell lines and at all concentrations tested (10 nM to 2 μM), +36 GFP delivered ~10- to 100-fold more mCherry than Tat or Arg_{10}. At concentrations ≤ 100 nM, +36 GFP also delivered ~6- to 20-fold more mCherry than penetratin, which approached the potency of +36 GFP only in HeLa cells and only at the highest tested concentrations (1 μM and 2 μM). These results suggest that +36 GFP is a significantly more potent protein transduction agent than the widely used Tat, Arg_{10}, and penetratin, especially at sub-micromolar concentrations. Delivery of mCherry into cells by +36 GFP was confirmed by live-cell confocal fluorescence

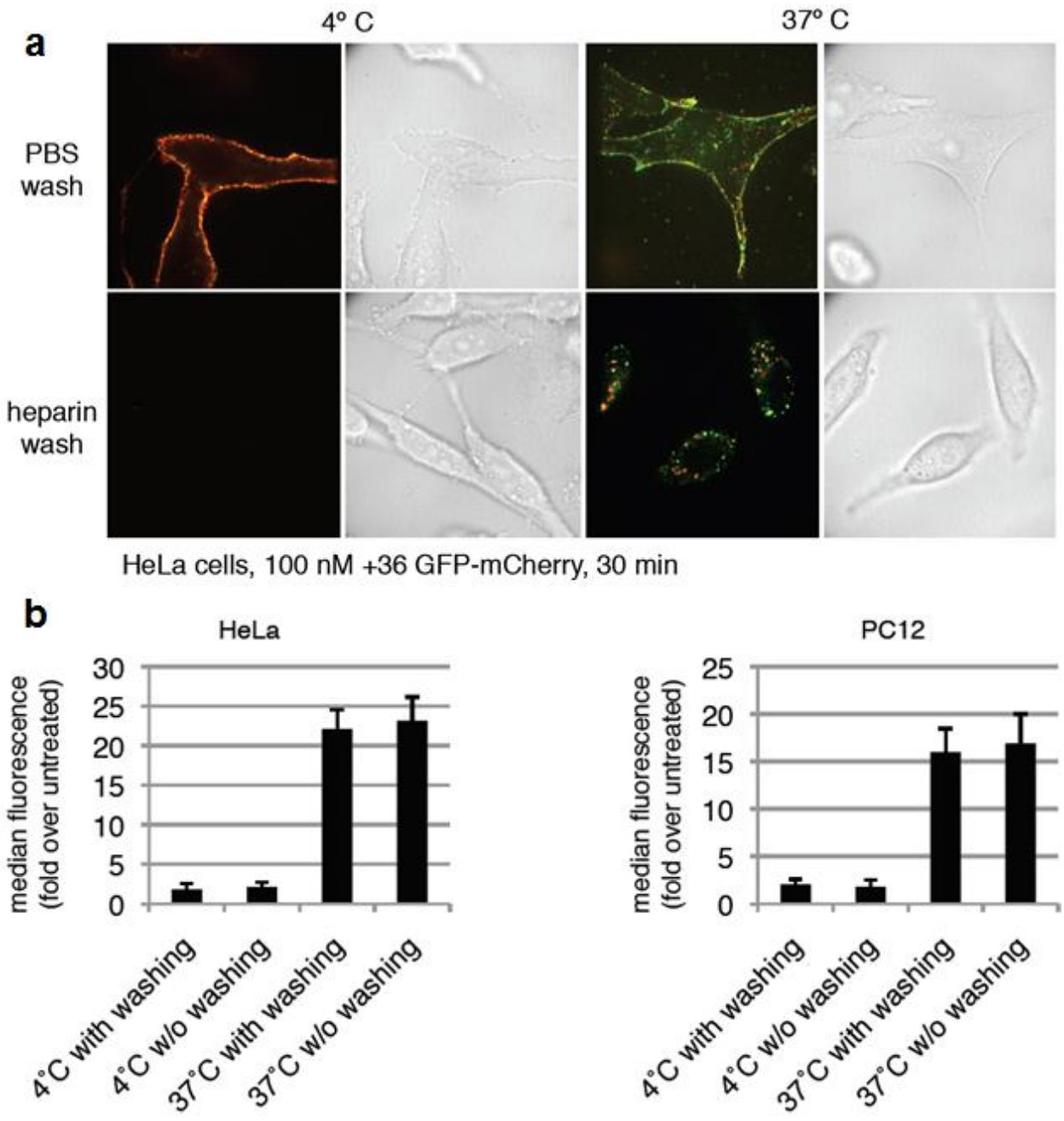

Figure 1.6 Membrane-bound protein is removed by heparin washing conditions. (a) Livecell fluorescence microscopy indicates that at 4 °C +36 GFP-mCherry is membrane-bound but not internalized. After washing with heparin (but not after washing with PBS), this +36 GFP-mCherry signal is largely removed. At 37 °C, most of +36 GFP-mCherry signal remains even after heparin washing, consistent with internalization of +36 GFP-mCherry. **(b)** HeLa and PC12 cells subjected to the conditions described in **(a**) were trypsinized (which destroys surface-bound mCherry) then analyzed by flow cytometry. Cells incubated with +36 GFP-mCherry at 4 °C do not show significant mCherry fluorescence compared to cells incubated at 37 °C, further suggesting that the signal at 37 °C represents internalized protein signal, and that internalization at 4 °C is inefficient.

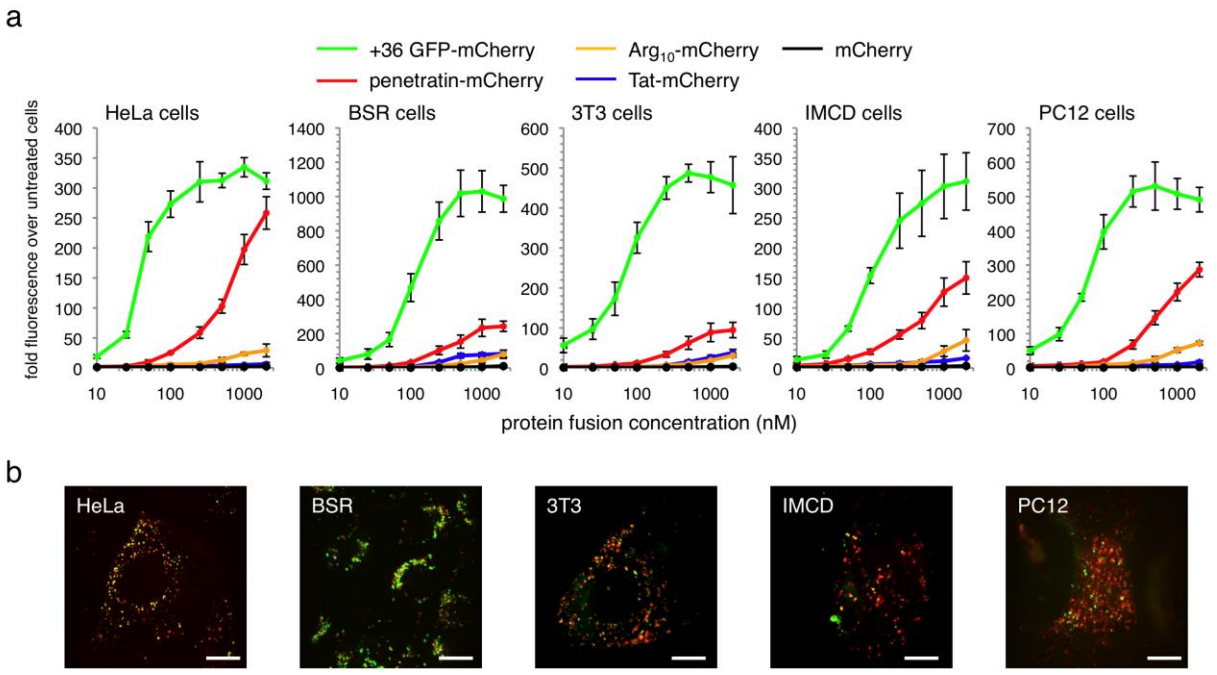

Figure 1.7 Comparison of mCherry delivery by +36 GFP, Tat, Arg_{10}, and penetratin. (**a**) Flow cytometry of HeLa, BSR, 3T3, PC12 and IMCD cells incubated in the presence of the specified concentrations of +36 GFP-mCherry, Tat-mCherry, Arg_{10}-mCherry, penetratin-mCherry or wild-type mCherry alone for 4 h at 37 °C. Cells were washed three times with 20 U/mL heparin in PBS to remove membrane-bound protein before analysis. Error bars represent the standard error of three independent biological replicates. (**b**) Confocal fluorescence microscopy of live cells incubated with 100 nM +36 GFP-mCherry for 4 h at 37 °C. Red color represents mCherry signal; green color represents +36 GFP signal. The scale bar is 15 μm.

microscopy (Figure 1.7b) and by comparison with control experiments in which endocytosis was blocked at 4˚C and protein remains surface-bound (Figure 1.6).

To study the ability of proteins delivered with +36 GFP to access the cytosol, we generated a ubiquitin-+36 GFP fusion in which the C-terminus of ubiquitin was directly followed by +36 GFP. A direct fusion of this type is recognized and processed by cytosolic deubiquitinases (DUBs), and DUB-dependent deubiquitination has previously been used as an indicator of cytosolic exposure.[12, 13] A mutant form of ubiquitin (G76V) that is not a substrate

for DUBs[12] was similarly fused to +36 GFP to distinguish the effect of cytosolic DUBs from non-specific proteolysis.

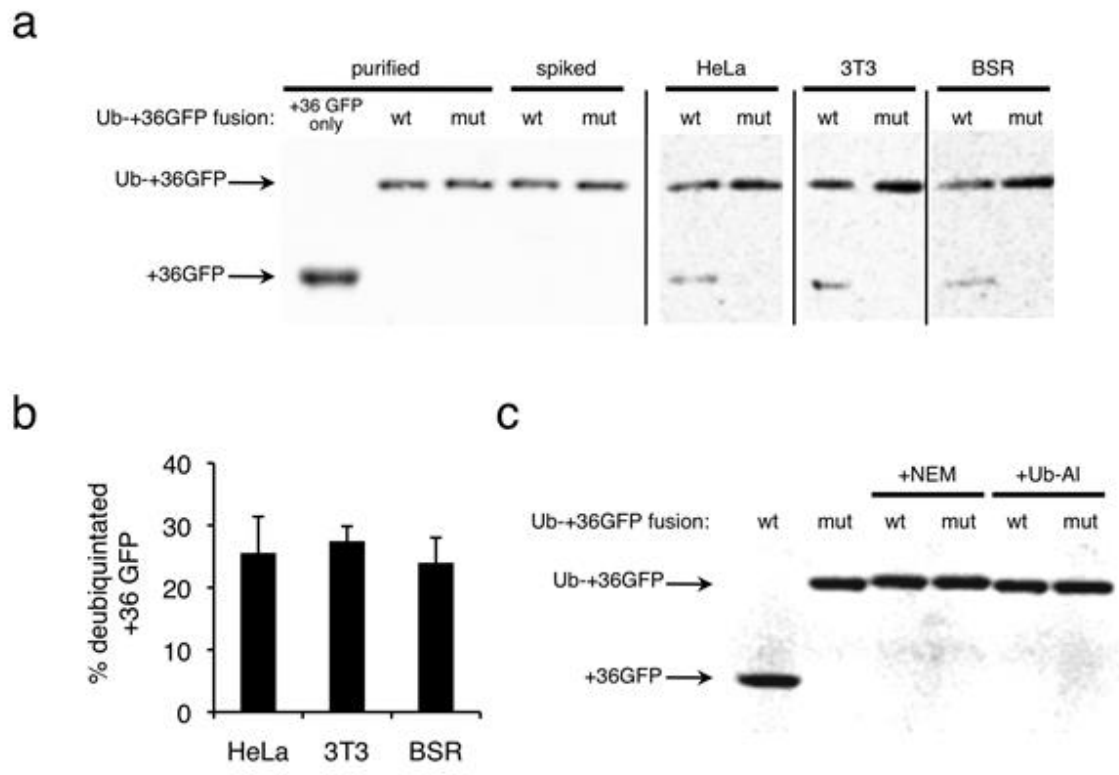

Figure 1.8 Deubiquitination suggests cytosolic exposure of a ubiquitin-+36 GFP fusion protein. (a) Western blots using anti-GFP antibodies. Lanes 1-3: purified protein samples of +36 GFP, wild-type ubiquitin–+36 GFP fusion (wt) or G76V mutant ubiquitin–+36 GFP fusion (mut). Lanes 4 and 5: purified protein spiked into HeLa cell lysate to confirm that lysis conditions do not affect fusion protein integrity. Lanes 6-11: the indicated cells were treated with 100 nM of either the wt or mutant ubiquitin–+36 GFP for 1 h, then lysed. (**b**) Mean extent of deubiquitination of wt ubiquitin–+36 GFP fusion protein in HeLa, 3T3, and BSR cells. Error bars reflect the standard deviation of three independent biological replicates. (**c**) *In vitro* deubiquitination control experiment. Ubiquitin–+36 GFP fusion proteins were incubated in either HeLa cytosolic extract or in HeLa cytosolic extract containing one of two DUB inhibitors, 10 mM *N*-ethylmaleimide (NEM) or 20 µg/mL ubiquitin-aldehyde (Ub-Al) for 1 h at 37 °C.

After a 1 h incubation of HeLa cells, 3T3, or BSR cells with either 100 nM ubiquitin-+36 GFP or 100 nM ubiquitin G76V +36 GFP, a significant fraction (HeLa: 25 ± 5.8%; 3T3: 27 ± 2.4%; BSR: 24 ± 4.0%) of internalized +36 GFP was deubiquitinated, producing a protein equal in size to +36 GFP (Figure 1.8a, b). In contrast, in all cases the G76V mutant-+36 GFP fusion

was not appreciably cleaved, indicating that this reduction in size does not arise from non-specific endosomal proteases but instead from the action of DUBs. Ubiquitin-+36 GFP spiked into the cell lysis buffer prior to harvesting untreated cells was not cleaved (Figure 1.8a), indicating that the observed deubiquitination is a result of exposure to cytosolic DUBs, and not due to contact with DUBs during the cell-harvesting procedure. Additionally, ubiquitin-+36 GFP was completely deubiquitinated when incubated in HeLa cytosolic extract for 1 h, while the DUB inhibitors *N*-ethylmaleimide[14] and ubiquitin-aldehyde[15] blocked deubiquitination (Figure 1.8c), further suggesting that the cleavage of ubiquitin-+36 GFP is a result of DUB activity. Collectively, these results demonstrate that some of the ubiquitin-+36 GFP protein fusion can access cytosolic enzymes in three distinct mammalian cell lines.

Next we compared the ability of +36 GFP, Tat, Arg_{10}, and penetratin to deliver a functional enzyme, Cre recombinase, into a variety of mammalian cells. Exogenously delivered Cre must escape the endosome, localize to the nucleus, and tetramerize to mediate DNA recombination.[16] We generated +36 GFP-Cre, Tat-Cre, Arg_{10}-Cre, and penetratin-Cre fusion proteins and tested their ability to effect recombination in HeLa cells transiently transfected with pCALNL-DsRed2,[17] a DsRed2-based Cre activity reporter plasmid. After incubation with 100-1000 nM of each Cre fusion protein for 4 h in serum-free media, cells were washed to remove surface bound protein and incubated in full media for 48 h. Delivery of Cre was assayed by following DsRed2 expression using flow cytometry and fluorescence microscopy (Figure 1.9a). We observed that +36 GFP-Cre generated ~2- to 5-fold more recombinants than the corresponding fusions with Tat, Arg_{10}, or penetratin.

Cre delivery was further evaluated in a NIH-3T3 cell line harboring an integrated lacZ-based Cre-reporter[5]. After incubation, treatment, and washing as described above, these cells

were stained with X-Gal to identify recombinants. Consistent with the HeLa cell results, +36 GFP-Cre resulted in more efficient generation of recombinants than Tat, Arg_{10}, or penetratin. The efficacy of +36 GFP-Cre was 10 to 100-fold higher than that of the other Cre fusions at 100 nM, 10-fold higher at 500 nM, and 5-fold more at 1 µM (Figure 1.9b). These findings together indicate that +36 GFP can deliver substantially more functional Cre than Tat, Arg_{10}, or penetratin in these cell lines.

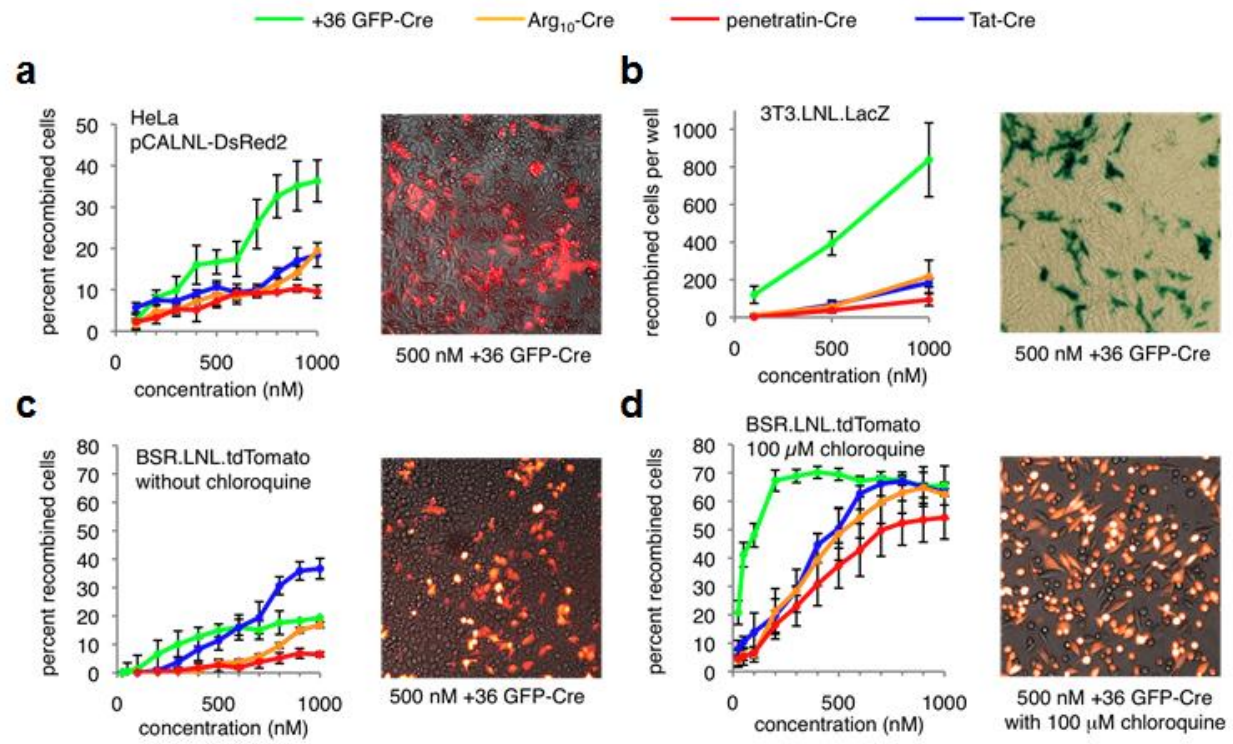

Figure 1.9 Delivery of active Cre recombinase into mammalian cells *in vitro*. (**a**) Cre-mediated recombination in HeLa cells transiently transfected with pCALNL-DsRed2 and treated with +36 GFP-Cre, Tat-Cre, Arg_{10}-Cre, or penetratin-Cre for 4 h at 37 °C. The image is an overlay of DsRed2 signal and brightfield images of HeLa cells transfected with pCALNL-DsRed2 and treated with 100 nM +36 GFP–Cre. (**b**) Cre-mediated recombination in 3T3.LNL.LacZ cells treated with +36 GFP-Cre, Tat-Cre, Arg_{10}-Cre, or penetratin-Cre for 4 h at 37 °C. The image is of 3T3.loxP.lacZ cells treated with 500 nM +36 GFP–Cre and stained with X-Gal. (**c**) Cre-mediated recombination in BSR.LNL.tdTomato cells treated with +36 GFP-Cre, Tat-Cre, Arg_{10}-Cre, or penetratin-Cre for 4 h at 37 °C. The image is an overlay of tdTomato signal and brightfield images of BSR.LNL.tdTomato cells treated with 100 nM +36 GFP–Cre. (**d**) Identical to (**c**) but with the addition of 100 µM chloroquine during and after protein treatment. In (**a**)-(**d**), error bars reflect the standard error of three independent biological replicates.

Next, we used BSR cells to generate a Cre reporter cell line conditionally expressing the tdTomato fluorescent protein after Cre-mediated DNA recombination. Following treatment as described above, Cre-mediated recombination was quantified by flow cytometry. In this cell line, +36 GFP was 2- to 15-fold more effective than Arg_{10} or penetratin (Figure 1.9c). At low concentrations, +36 GFP delivered modestly higher levels of functional Cre than Tat, while higher concentrations Tat-Cre generated ~2-fold more recombinants than +36 GFP-Cre. This cell line exhibits unusual features, including a high metabolic rate (doubling time = ~12 h[18]), that led us to speculate that endosomal trafficking to lysosomes may be unusually efficient in these BSR cells compared with HeLa- and 3T3-based reporter cells. Indeed, when BSR cells were incubated with the Cre fusions in the presence of 100 μM chloroquine, an inhibitor of protein lysosomal degradation,[19] we observed a dramatic increase in the number of recombinants arising from +36 GFP-Cre treatment and modest improvements for Tat-Cre, Arg_{10}-Cre, and penetratin-Cre, such that at all concentrations tested +36 GFP-Cre delivered more recombinase activity than any of the other proteins (Figure 1.9d). These results suggest that the unique cell-penetration potency of +36 GFP can be more fully exploited by extending the time available for internalized protein to escape endosomes, resulting in even higher levels of delivered functional protein. Supporting this hypothesis, *in vitro* recombinase assays show that while +36 GFP-Cre is not active as the full length fusion, at pH 5.5-6.5 cathepsin B, a ubiquitous mammalian endosomal protease, will cleave +36 GFP-Cre to reveal +36 GFP and active Cre (Figure 1.10a). At pH 5.0, the exopeptidase activity of cathepsin B is maximized [20], and we find no full length Cre remaining from the cleavage reaction nor do we observe recombinase activity *in vitro* (Figure 1.10b, c). This is consistent with choloroquine's mechanism for inhibiting lysosomal protein degradation, which is by preventing complete acidification of lysosomes[21].

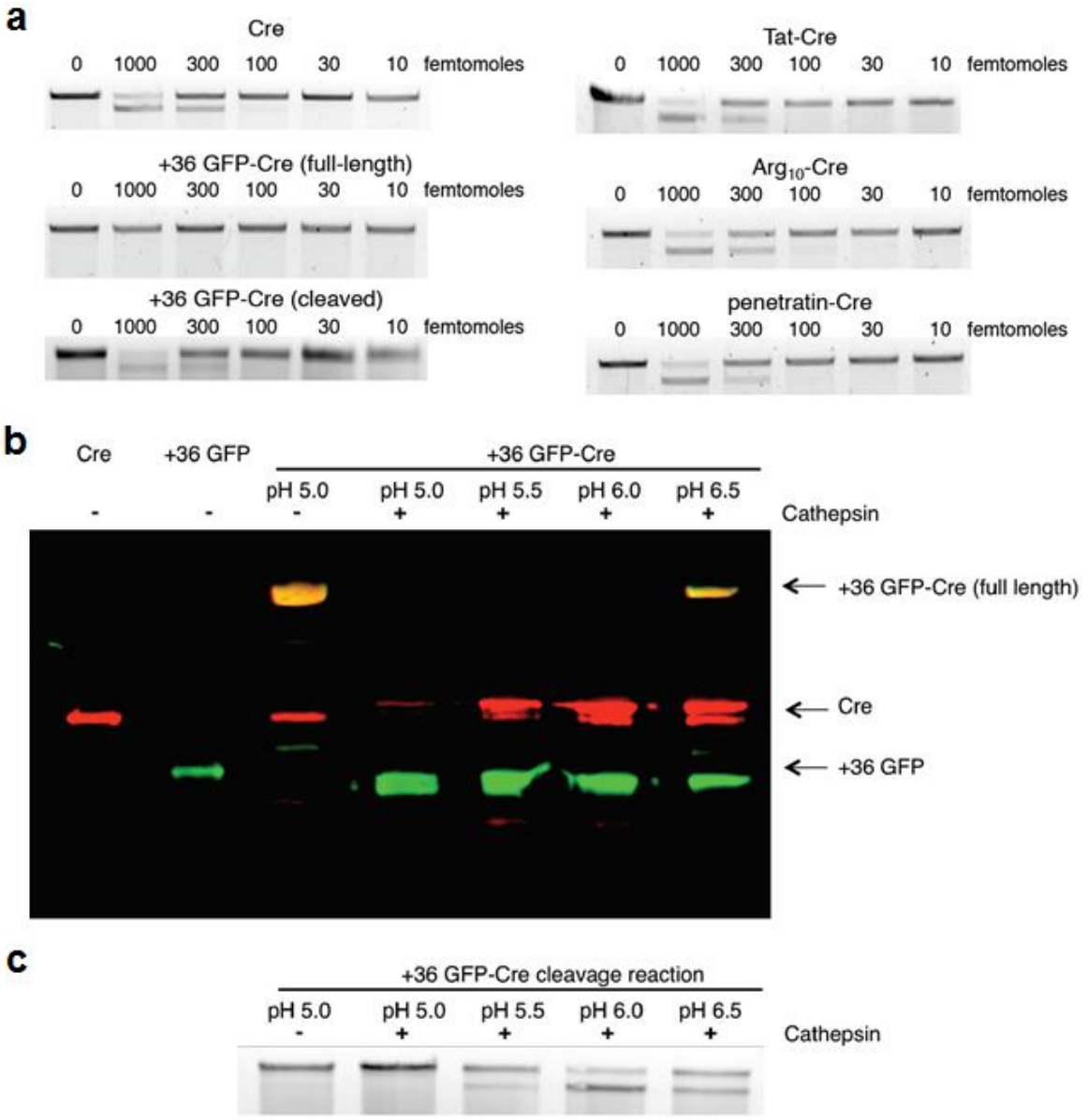

Figure 1.10 Protease cleavage dependence of +36 GFP-Cre activity. (**a**) The concentration of Cre proteins was verified explicitly by in vitro activity assay. All Cre proteins showed roughly 50% cleavage activity at 300 femtomoles/reaction and nearly quantitative activity at 1000 femtomoles/reaction. Note that +36 GFP-Cre did not exhibit activity as the fusion protein but exhibited wild-type-like activity once cleaved by cathepsin B. Cleavage and activity of +36 GFP-Cre incubated with cathepsin B in buffer at various pH values. (**b**) Western blot of +36 GFP-Cre incubated with or without cathepsin B. Incubation of 50 picomoles of +36 GFP-Cre with or without 0.5 µg cathepsin B (Sigma) in 20 µL of 50 mM MES, pH 5.0-6.5 at 37º C for 45 min results in variable amounts of full-length +36 GFP-Cre, Cre, and +36 GFP. Note that at pH 5.0, Cre is largely degraded by cathepsin B, while at higher pH values, Cre is cleaved from +36 GFP but is not degraded to a significant extent. (c) The in vitro recombinase activity of the products of +36 GFP-Cre incubation with cathepsin B at the pH values indicated demonstrate that +36 GFP-Cre is not active as the full-length fusion or after incubation (degradation) with cathepsin B at pH 5.0. The upper band is Cre substrate DNA, while the lower band is recombination product, as in (a).

Finally, we tested +36 GFP as a protein delivery agent *in vivo*. First we examined the tissue penetration of +36 GFP in the adult mouse retina. We injected 0.5 µL of 100 µM +36 GFP into the subretinal space of CD1 adult mice. After 6 h, the retinas were harvested and analyzed by fluorescence microscopy (Figure 1.11a). Most of +36 GFP was observed by the photoreceptor outer segments, but significant signal was observed throughout the retina, including all three nuclear layers (the outer, inner, and ganglion cell layers) as well as in the cell processes.

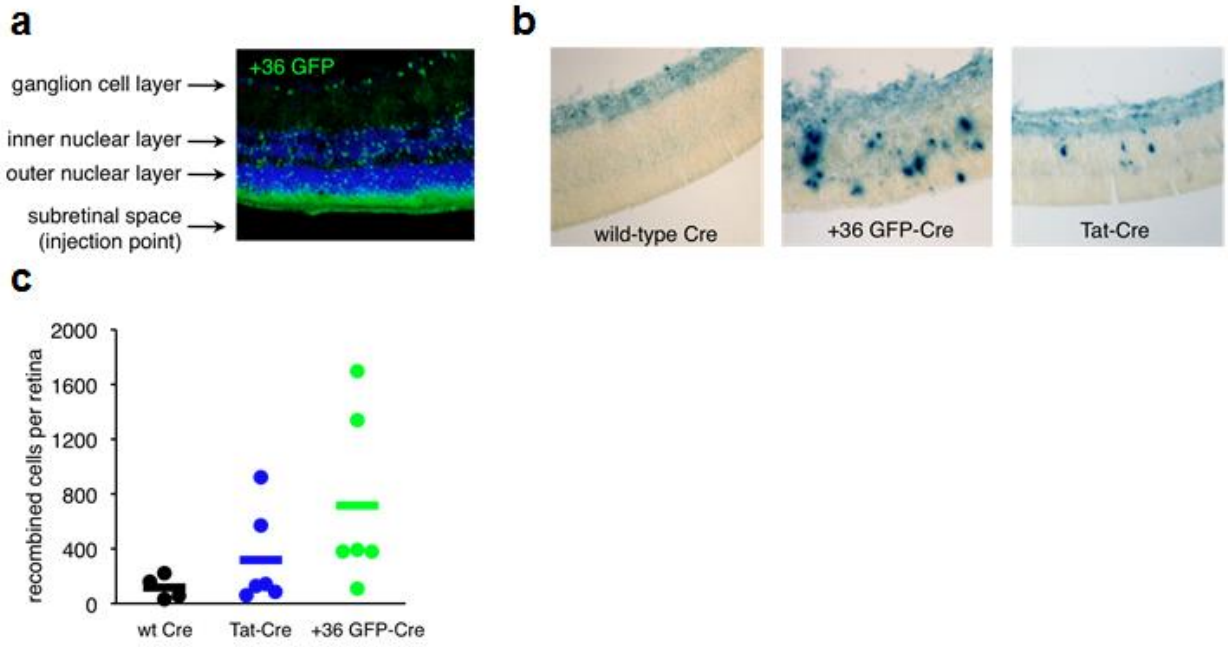

Figure 1.11 Delivery of active Cre recombinase into mouse retinal cells *in vivo*. (a) Fluorescence microscopy of a retinal section of a CD1 adult mouse injected with 0.5 µL of 100 µM +36 GFP. The retina was harvested and analyzed six h after injection. GFP fluorescence is shown in green and DAPI nuclear stain is shown in blue. **(b)** Retinal sections of neonatal RC::PFwe mouse pups harboring a nuclear LacZ reporter of Cre activity. Three days after injection of 0.5 µL of 40 µM wild-type Cre, Tat-Cre, or +36 GFP-Cre, retinae were harvested, fixed, and stained with X-gal. **(c)** Dots on the graph represent the total number of recombined cells counted in each retina. The horizontal bar represents the average number of recombined cells per retina for each protein injected (n = 4 for wild-type Cre, n = 6 for Tat-Cre, n = 6 for +36 GFP-Cre).

To test ability of +36 GFP to deliver functional protein *in vivo*, we injected +36 GFP-Cre into the subretinal space of RC::PFwe mouse p0 pups containing a LoxP-flanked transcriptional

terminator upstream of a nuclear lacZ reporter gene.[22] Three days after injection of 0.5 μL of 40 μM wild-type Cre, Tat-Cre, or +36 GFP-Cre, retinae were harvested, fixed, and stained with X-gal (Figure 1.11b). Injection of +36 GFP-Cre generated an average of 715 recombined cells per injected retina (n = 6), Tat-Cre generated an average of 318 recombined cells (n = 6) while wild-type Cre generated an average of 117 recombined cells per retina (n = 4) (Figure 1.11c). To our knowledge, this is the first report of functional delivery of an enzyme into retinal cells *in vivo.*

Discussion

Side-by-side comparisons of +36 GFP, Tat, Arg_{10} and penetratin fused to mCherry or Cre recombinase revealed that fusions with supercharged GFP result in dramatically higher levels of internalized protein (up to ~100-fold) and in significantly greater efficiencies of Cre-induced recombination (up to ~10-fold) than three currently used protein transduction domains in a variety of mammalian cell lines. Functional Cre recombinase can also be delivered to cells upon injection *in vivo* using +36 GFP. These results collectively demonstrate the potential of supercharged proteins as an unusually potent *in vitro* and *in vivo* protein delivery platform.

Methods

Live-Cell Imaging

Cells were plated onto 35 mm glass-bottom microwell dishes with a No. 1.5 cover glass (MatTek) at a density of 10^6 cells per plate. After 18 h, cells were washed once with cold PBS and incubated with protein in serum-free DMEM. After incubation, cells were washed three times with cold 20 U/mL heparin in PBS to remove membrane-bound protein, imaged in prewarmed HEPES imaging solution (1 mM $MgCl_2$, 5 mM KCl, 5 mM $CaCl_2$, 150 mM NaCl, 1.9 g/L glucose, 1.9 g/L albumin, 20 mM HEPES, pH 7.4, 37 °C). Cells were imaged on an

Olympus IX71 spinning disk confocal microscope on a heated stage with a 100X objective lens. GFP and mCherry were visualized by with a 491 nm and 561 nm excitation laser, respectively. Images were prepared using OpenLab software.

For live-cell images of Cre reporter cells, cells were treated with 500 nM +36 GFP-Cre as described below for analysis of Cre recombination and imaged on an Olympus IX51 fluorescent microscope with a DP30BW black and white camera. Images are false-color overlays of the fluorescent signal on a bright-field image. LacZ-positive cells were imaged on an IX70 microscope under bright-field illumination with a DP70 camera.

Flow Cytometry and mCherry Delivery Assays

Cells were plated into a 48-well plate at a density of 5 x 10^5 cells per well. After 18 h, cells were washed once with cold PBS and incubated with protein in serum-free DMEM. After incubation, cells were washed three times with cold 20 U/mL heparin in PBS to remove membrane-bound protein, trypsinized, resuspended in 500 µL of full media and placed on ice. Cells were analyzed on either a LSRII or Fortessa flow cytometer (BD Biosciences) for GFP internalization (ex: 488 nm) or mCherry internalization (ex: 561 nm). Cells were gated for live cells and at least 5 x 10^4 live cells were analyzed for each treatment. Data was analyzed with FlowJo software (Tree Star, Inc.)

In Vitro Cre Delivery Assays

Hela cells were plated at 3 x 10^4 cells/well in 48-well plates. After 16 h, cells were transfected with pCALNL-DsRed2[17] using Effectene transfection reagent (Qiagen). After incubation with 100-1000 nM of each Cre fusion protein for 4 h in serum-free DMEM, cells were washed three times with 20 U/mL heparin in PBS and incubated in full media for 48 h.

Delivery of Cre was assayed by following DsRed2 expression using flow cytometry and fluorescence microscopy.

Cre reporter 3T3 cells were plated at 1 x 10^5 cells/well in 48-well plates. After 16 h, cells were incubated with various concentrations of protein for 4 h in serum-free media. Cells were washed with three times with 20 U/mL heparin in PBS and incubated in full media for 48 h. Recombined cells were quantified by X-gal staining and manual counting.

BSR cells were obtained from Matthias Schnell (Thomas Jefferson University). A pQCXIX MMLV retrovirus (Clontech) containing the tdTomato cre reporter construct was generated by subcloning the tdTomato gene (Clontech) into a pCALNL backbone[17] and packaged using Plat-E cells[23]. BSR cells were infected with retrovirus and integrants were selected for one week in the presence of 1 mg/ml G418 (Sigma). BSR.LNL.tdTomato cells were plated at 1 x 10^5 cells/well in 48 well plates. After 16 h, cells were incubated with various concentrations of protein for 4 h in serum-free media. Cells were washed with three times with 20 U/mL heparin in PBS and incubated in full media for 48 h.

For chloroquine treatment of BSR cells, cells were incubated with Cre fusion proteins for 4 h in serum-free media containing 100 μM chloroquine, washed three times with 20 U/mL heparin in PBS, and incubated 12 h in full media containing 100 μM chloroquine. Following this incubation, cells were washed once with PBS and incubated a further 36 h in full media without chloroquine. Delivery of Cre was assayed by following tdTomato expression using flow cytometry and fluorescence microscopy.

For fluorescent Cre reporters, recombinants were identified by flow cytometry as those cells of the live-cell population that exhibited significantly higher fluorescence than that of non-

treated reporter cells. Typically, the recombined population exhibited fluorescence at least 10-fold higher than the non-recombined cells and were readily detected as a distinct subpopulation. Fluorescence gates were drawn accordingly to quantitate recombined and non-recombined cells.

In Vivo Cre Delivery

RC::PFwe mice were obtained from Susan Dymecki (Harvard University). All of the experiments in this study were approved by the Institutional Animal Care and Use Committee at Harvard University. Adult CD1 mice were subretinally injected with 0.5 μL of 100 μM +36 GFP. After 6 h, the retinas were harvested and analyzed by fluorescence microscopy. p0 pups were subretinally injected with 0.5 μL of 40 μM wtCre, Tat-Cre, or +36 GFP-Cre. After 72 h, retinae were harvested and fixed with 0.5% glutaraldehyde. Fixed retinae were stained with X-gal overnight and embedded in 50% OCT, 50% of 30% sucrose and stored at -80° C. Retinae were cut into 30 μm sections and imaged for X-gal staining on a Zeiss Axiophot brightfield microscope with a Nikon CXM-1200F camera. Delivery of Cre was assayed by manually counting $LacZ^+$ cells.

References

1. Zelphati, O.; Wang, Y.; Kitada, S.; Reed, J. C.; Felgner, P. L.; Corbeil, J., Intracellular delivery of proteins with a new lipid-mediated delivery system. *J. Biol. Chem.* **2001,** *276* (37), 35103-35110.
2. Hasadsri, L.; Kreuter, J.; Hattori, H.; Iwasaki, T.; George, J. M., Functional protein delivery into neurons using polymeric nanoparticles. *J. Biol. Chem.* **2009**.
3. Gabel, C. A.; Foster, S. A., Mannose 6-phosphate receptor-mediated endocytosis of acid hydrolases: internalization of beta-glucuronidase is accompanied by a limited dephosphorylation. *J. Cell Biol.* **1986,** *103* (5), 1817-1827.

4. Rizk, S. S.; Luchniak, A.; Uysal, S.; Brawley, C. M.; Rock, R. S.; Kossiakoff, A. A., An engineered substance P variant for receptor-mediated delivery of synthetic antibodies into tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106* (27), 11011-11015.

5. Wadia, J. S.; Dowdy, S. F., Modulation of cellular function by TAT mediated transduction of full length proteins. *Curr. Protein Pept. Sci.* **2003,** *4* (2), 97-104.

6. Heitz, F.; Morris, M. C.; Divita, G., Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. *British journal of pharmacology* **2009,** *157* (2), 195-206.

7. Cai, S. R.; Xu, G.; Becker-Hapak, M.; Ma, M.; Dowdy, S. F.; McLeod, H. L., The kinetics and tissue distribution of protein transduction in mice. *Eur. J. Pharm. Sci.* **2006,** *27* (4), 311-319.

8. Caron, N. J.; Torrente, Y.; Camirand, G.; Bujold, M.; Chapdelaine, P.; Leriche, K.; Bresolin, N.; Tremblay, J. P., Intracellular delivery of a Tat-eGFP fusion protein into muscle cells. **2001,** *3* (3), 310-318.

9. Lawrence, M. S.; Phillips, K. J.; Liu, D. R., Supercharging proteins can impart unusual resilience. **2007,** *129* (33), 10110-10112.

10. McNaughton, B. R.; Cronican, J. J.; Thompson, D. B.; Liu, D. R., Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106* (15), 6111-6116.

11. Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N.; Palmer, A. E.; Tsien, R. Y., Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat. Biotechnol.* **2004,** *22* (12), 1567-1572.

12. Loison, F.; Nizard, P.; Sourisseau, T.; Le Goff, P.; Debure, L.; Le Drean, Y.; Michel, D., A ubiquitin-based assay for the cytosolic uptake of protein transduction domains. *Mol. Ther.* **2005,** *11* (2), 205-214.

13. Varshavsky, A., Ubiquitin fusion technique and related methods. *Methods Enzymol.* **2005,** *399*, 777-799.

14. Borodovsky, A.; Kessler, B. M.; Casagrande, R.; Overkleeft, H. S.; Wilkinson, K. D.; Ploegh, H. L., A novel active site-directed probe specific for deubiquitylating enzymes reveals proteasome association of USP14. *EMBO J.* **2001,** *20* (18), 5187-5196.

15. Hershko, A.; Rose, I. A., Ubiquitin-aldehyde: a general inhibitor of ubiquitin-recycling processes. *Proc Natl Acad Sci U S A* **1987,** *84* (7), 1829-33.

16. Guo, F.; Gopaul, D. N.; van Duyne, G. D., Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **1997,** *389* (6646), 40-46.

17. Matsuda, T.; Cepko, C. L., Controlled expression of transgenes introduced by in vivo electroporation. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (3), 1027-1032.

18. Chang, L. S.; Pater, M. M.; Hutchinson, N. I.; di Mayorca, G., Transformation by purified early genes of simian virus 40. *Virology* **1984,** *133* (2), 341-353.

19. Seglen, P. O.; Grinde, B.; Solheim, A. E., Inhibition of the lysosomal pathway of protein degradation in isolated rat hepatocytes by ammonia, methylamine, chloroquine and leupeptin. *Eur. J. Biochem.* **1979,** *95* (2), 215-225.

20. Almeida, P. C.; Nantes, I. L.; Chagas, J. R.; Rizzi, C. C.; Faljoni-Alario, A.; Carmona, E.; Juliano, L.; Nader, H. B.; Tersariol, I. L., Cathepsin B activity regulation. Heparin-like glycosaminogylcans protect human cathepsin B from alkaline pH-induced inactivation. *J. Biol. Chem.* **2001,** *276* (2), 944-51.

21. Anderson, R. G.; Falck, J. R.; Goldstein, J. L.; Brown, M. S., Visualization of acidic organelles in i ntact cells by electron microscopy. *Proc Natl Acad Sci U S A* **1984,** *81* (15), 4838-42.

22. Farago, A. F.; Awatramani, R. B.; Dymecki, S. M., Assembly of the brainstem cochlear nuclear complex is revealed by intersectional and subtractive genetic fate maps. *Neuron* **2006,** *50* (2), 205-218.

23. Morita, S.; Kojima, T.; Kitamura, T., Plat-E: an efficient and stable system for transient packaging of retroviruses. **2000,** *7* (12), 1063-1066.

Chapter 2:

Cellular Uptake Mechanisms and Endosomal Trafficking of Supercharged Proteins and Liposomal Delivery of Anionic Supercharged Protein Fusions

Abstract

Supercharged proteins can deliver functional macromolecules into the cytoplasm of mammalian cells with potencies that exceed those of cationic peptides. The structural features of supercharged proteins that determine their delivery effectiveness and the intracellular fate of supercharged proteins once they enter cells have not yet been studied. Using a large set of supercharged GFP (scGFP) variants, we found that the level of cellular uptake is sigmoidally related to net charge, and that scGFPs enter cells through multiple pathways including clathrin-dependent endocytosis and macropinocytosis. Supercharged proteins activate Rho and ERK1/2, and also alter the endocytic transport of transferrin and EGF. Finally, we discovered that the intracellular trafficking of endosomes containing scGFPs is altered in a manner that correlates with protein delivery potency. Collectively, our findings establish basic structure-activity relationships of supercharged proteins and implicate the modulation of endosomal trafficking as a determinant of cell-penetration and macromolecule-delivery efficiency.

Introduction

The vast majority of nucleic acids and proteins encoded by the human genome are intracellular. Molecular strategies to perturb the function of most biological targets for research or therapeutic purposes therefore require agents that can enter cells. While membrane-permeable small molecules have dominated therapeutics over the past several decades, the use of macromolecules to address biomedical targets has more recently become a focus of intense research[1], resulting in a number of macromolecular human drugs[2]. Macromolecules can offer significant advantages over traditional small molecule-based therapeutics. Macromolecules possess sizes and folding energies that are ideal for catalyzing chemical reactions, potently and selectively binding to extended target surface areas, and encoding gene products. These key

features, juxtaposed with the inability of virtually all macromolecules to spontaneously enter cells, create an urgent need to develop effective and general methods for the delivery of functional proteins and nucleic acids into mammalian cells[3,4].

While a host of protein delivery methods have been developed over the past decade—most notably those based on cationic cell-penetrating peptides (CPPs)[5, 6,] but also including antibodies[7], receptor ligands[8], nanoparticles[9], and virus-like particles[10] — these approaches have not achieved widespread use, and modestly successful protein delivery has historically required high doses of purified protein. We recently developed a platform for macromolecule delivery *in vitro* and *in vivo* based on supercharged proteins (SCPs)[11, 12, 13]. SCPs possess extremely high positive theoretical net charge at their surfaces and candidate SCPs can be generated computationally from native, non-supercharged proteins[14]. Proteins can be mutated into SCPs without necessarily abolishing the protein's native structural or functional properties[14]. Recently, we identified a class of naturally occurring supercharged human proteins with theoretical net charge:molecular weight ratios similar to those of engineered SCPs[11]. These naturally occurring human SCPs exhibit similar potent cellular uptake and macromolecule-delivery properties as engineered SCPs. When compared to the most commonly used cell-penetrating peptides (CPPs) and commercial nucleic acid delivery reagents, SCPs can result in more effective protein and nucleic acid delivery across a range of cell and tissue types *in vitro* and *in vivo*[12,13,11].

The molecular mechanisms by which SCPs enter and are trafficked in cells are largely unknown. Previous studies suggest that cell entry of SCPs shares some mechanistic features with that of other cationic delivery reagents[13], such as binding to sulfated proteoglycans to mediate initial cellular association, followed by endocytosis and some degree of endosomal

escape to the cytosol. However, SCPs can be more efficient at achieving both cell entry and functional macromolecule delivery to the cytosol than CPPs[12,13]. It is not understood whether the high delivery potency of SCPs compared with CPPs results solely from the higher theoretical net charges attained by SCPs, or from the globular, structured nature of SCPs compared with much shorter and less structured peptide tags. Observations that support the latter hypothesis include: (*i*) increasing the theoretical net charge of CPPs beyond approximately +15 typically does not increase, and eventually decreases, their cell penetration potency[15]; (*ii*) substitution of arginines on the surfaces of small proteins has been shown to endow these proteins with cell-penetrating abilities[16] beyond those of similarly charged cationic peptides; (*iii*) the cell-penetration capabilities of oligoarginine peptides have been enhanced by the introduction of disulfide bonds that presumably stabilize the three-dimensional structure of the cationic peptide[17]; (*iv*) introducing spacer sequences between cationic residues throughout an oligoarginine peptide has also been shown to increase the effectiveness of cellular uptake[18]; (*v*) the distribution of cationic residues on the surface of proteins, and not merely charge magnitude, has been shown to alter the effectiveness of certain cell-penetrating proteins[19]. Consistent with these findings, we have observed that SCPs exhibit more potent cell penetration and nucleic acid delivery abilities than synthetic peptides of similar or even greater positive theoretical net charge[13].

In this work we investigate in detail the role of theoretical net charge and charge distribution on the cellular uptake and delivery properties of SCPs using supercharged GFP (scGFP) as a model. We examined the extent to which SCPs and CPPs alter endocytic processes using a wide range of biochemical assays and high-throughput confocal microscopy. We find that cellular uptake and delivery potencies are strongly charge-dependent, and that supercharged proteins outperform unstructured peptides of similar charge. We further find that delivery

potency correlates with the activation of specific cellular signaling markers, altered endocytosis, and subcellular trafficking. These findings provide mechanistic insights into the unusual ability of supercharged proteins to enter cells and deliver macromolecules, and inform the future development and application of these agents.

Results

Previously we observed that the CPPs Tat, Arg_{10}, and penetratin delivered protein into cells with up to 100-fold lower potencies than that of +36 GFP[12], and that protein and nucleic acid delivery efficiencies mediated by three supercharged GFP variants (+15, +25, and +36 GFP) correlated with their theoretical net charge[13]. These observations prompted us to characterize in detail (*i*) the relationship between net theoretical charge and cellular uptake potency, and (*ii*) the effect of charge distribution, either across the surface of a protein or within an unstructured peptide tag, on cellular uptake and delivery potency.

To elucidate these relationships, we generated a collection of 28 scGFP variants with a wide range of theoretical net charges as well as different patterns of charge distribution (**Figure 2.1**). Genes encoding the variants were constructed by combining DNA fragments derived from the coding sequences of previously developed GFPs: stGFP (starting GFP, a non-supercharged variant), +15 GFP, +25 GFP, +36 GFP, and +48 GFP[14]. Each resulting scGFP variant was expressed and purified from *E. coli* cells. The absorbance and fluorescence emission spectra of all 28 scGFPs were very similar (**Figure 2.2**), enabling us to use fluorescence to directly compare the cellular uptake of these proteins.

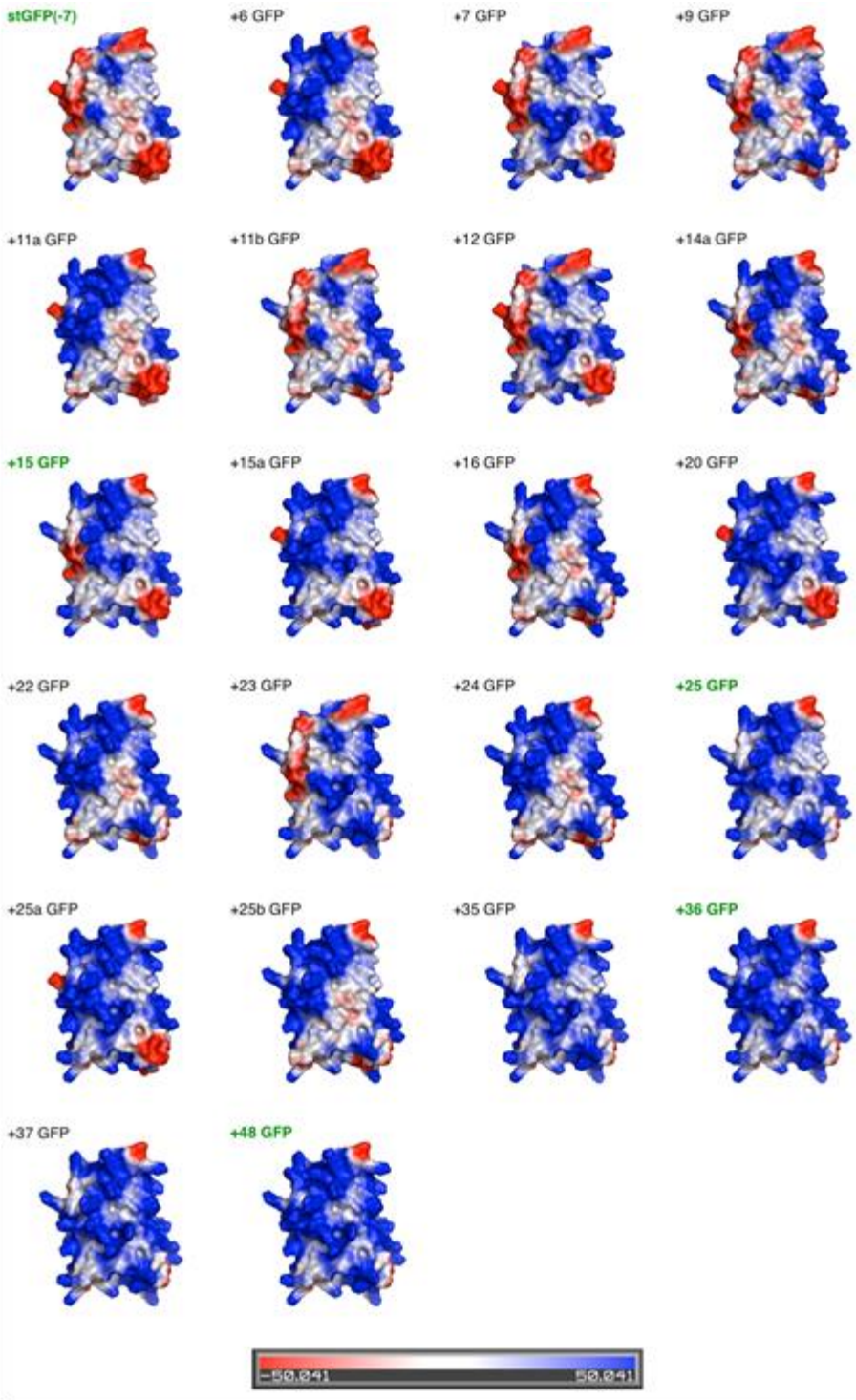

Figure 2.1 Electrostatic potential of the scGFPs. PyMol renderings of each scGFP variant in the same spatial orientation. The surfaces are colored from red (-) to blue (+) based on electrostatic potential to help visualize the difference between specific scGFP variants. The original scGFPs used to elaborate the rest are labeled in bold green text. The scale is in units of k_bT.

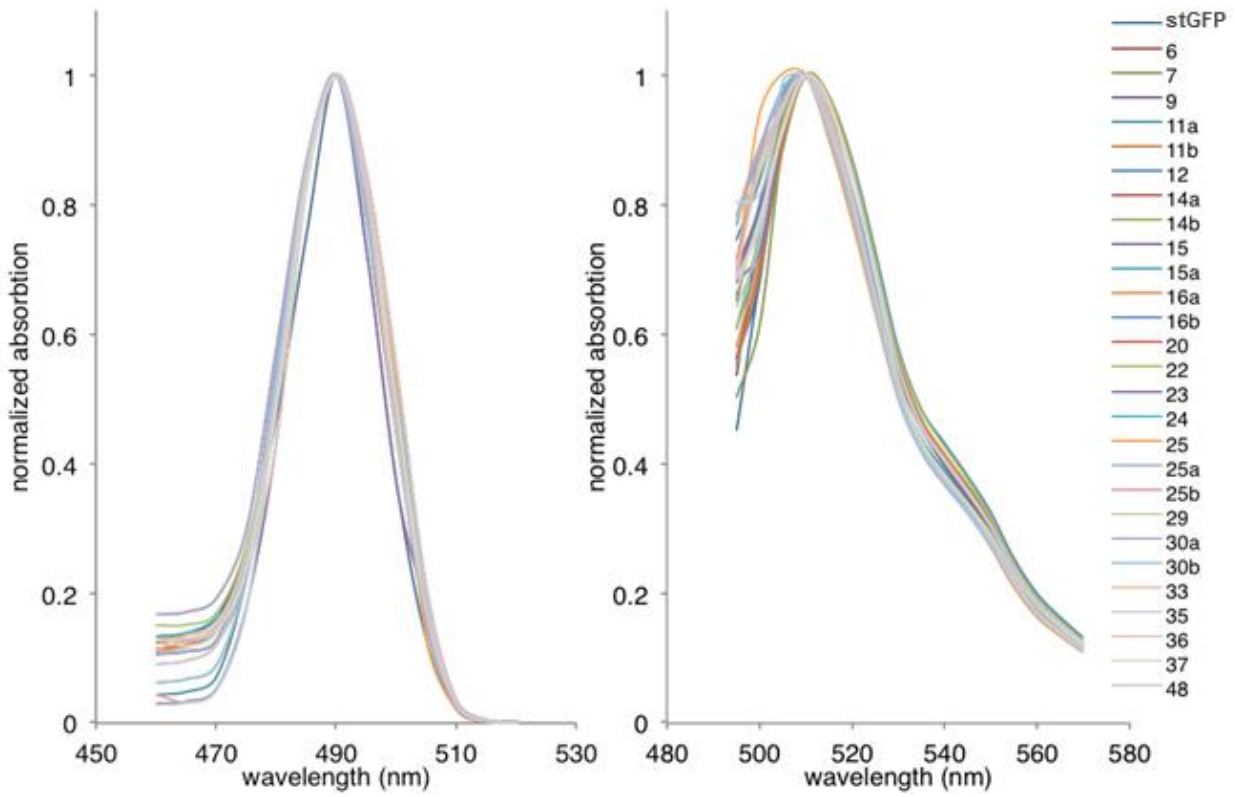

Figure 2.2 The absorbance and fluorescence emission spectra of scGFPs. 2 μM of each scGFP protein in PBS was loaded into a black clear-bottomed 96-well plate and scanned for absorbance (left) from 460 nm to 520 nm in 5 nm steps, and scanned for fluorescence emission with excitation at 470 nm from 495 nm to 570 nm in 5 nm steps (right).

To determine the relationship of charge to cell uptake, each scGFP variant was incubated with HeLa cells for 4 hours across a range of concentrations. Following incubation, cells were washed using conditions previously shown to remove excess non-internalized protein[12,13,11], trypsinized, and assayed for protein uptake by flow cytometry. For comparison, Tat-stGFP and a GFP engineered by Raines and coworkers to contain an arginine patch[20] were also incubated with HeLa cells and processed in an identical manner. The degree of cellular uptake was measured as the median cellular fluorescence of the population of cells. The relationship of theoretical net charge to cellular uptake was found to be strongly sigmoidal, and consistent

across doses ranging from 20 nM to 2 μM (**Figure 2.3**). Low-potency uptake, comparable to that of Tat peptide, was observed among proteins with net theoretical charge < ~+20.

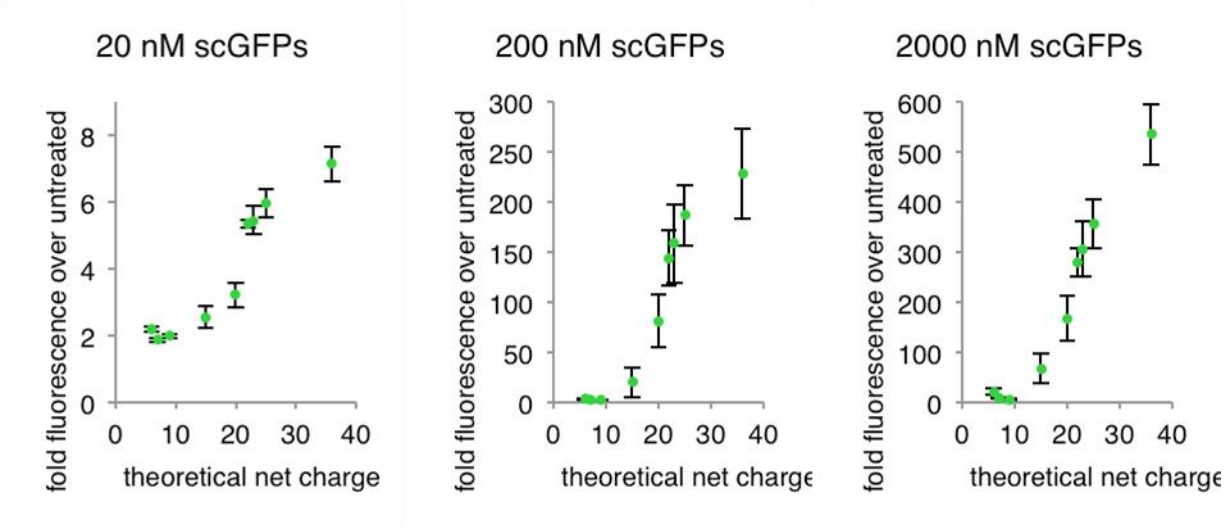

Figure 2.3 The charge-uptake relationship of scGFPs is consistent across protein dosage. HeLa cells treated with either 20 nM, 200 nM, or 2000 nM of each scGFP. The scGFPs included are +6, +7, +9, +15, +20, +22, +23, +25, and +36 GFP. Error bars represent the standard deviation of three independent replicates.

In contrast, a high-potency regime exhibiting 100-fold higher cellular uptake potency was observed for highly charged proteins (**Figure 2.4a**). For GFP, the inflection point between the low- and high-potency regimes occurs at a theoretical net charge of approximately +22 (**Figure 2.4a**, green points), well above the net charge of CPPs. As this net charge-to-cellular uptake relationship predicts, Tat-stGFP (+1 theoretical net charge) exhibited potency comparable to that of other modestly charged, low-potency charged GFP variants, including the Arg-grafted GFP[20] (**Figure 2.4a**). The striking change in cell penetration effectiveness from low- to high-potency variants suggests that these variants may enter cells through the involvement of different cellular interactions or through distinct uptake pathways.

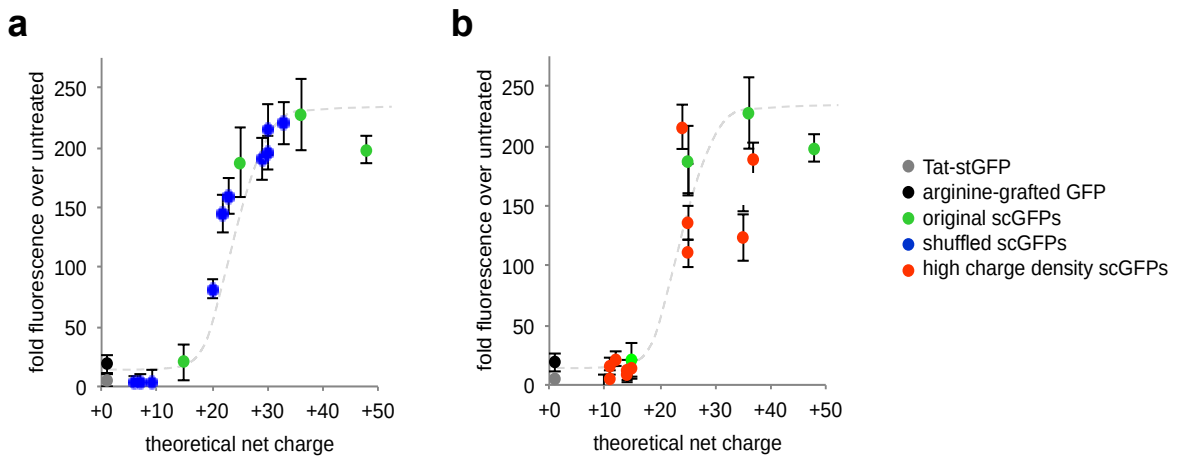

Figure 2.4 Charge dependence of cellular uptake supercharged GFPs. (**a**) and (**b**) HeLa cells were treated with 200 nM of each purified scGFP protein for 4 hours, washed to remove surface-bound protein and analyzed by flow cytometry. Plots show fold median GFP fluorescence intensity relative to untreated control cells. Blue points represent scGFPs generated by shuffling stGFP, +15 GFP, +25 GFP, and +36 GFP sequences. Red points represent scGFPs generated by shuffling stGFP, +15 GFP, +25 GFP, +36 GFP, and +48 GFP sequences to create proteins with less even charge density. Green points represent starting +15, +25, +36, and +48 GFPs. The grey point is Tat-stGFP, and the black point is an arginine-grafted GFP.

To examine the effect of SCP charge distribution on cell penetration, we generated a series of scGFPs that alter charge distribution without varying theoretical net charge, which was maintained at approximately +25 and +36. The variants within this series were constructed from segments of +48 GFP together with segments of +15 GFP and +25 GFP and therefore contained a more uneven cationic charge distribution than canonical +25 or +36 GFPs, with basic residues concentrated into more limited regions of the protein surface. We observed significantly different levels of cellular uptake potencies among scGFPs with similar or identical theoretical net charges but different charge distribution patterns (**Figure 2.4b**, red points). These results confirm that the distribution of cationic groups on the surface of a supercharged protein,

and not simply the theoretical net charge of the protein, contributes to its cellular uptake properties.

To account for the possibility that interaction between densely substituted cationic residues influences their protonation state and the net charge of scGFPs, we performed cation exchange and found that all proteins eluted in a manner consistent with their relative theoretical net charge rather than with their observed level of cellular uptake (**Figure 2.5a**). Additionally, we tested whether the voltage of the scGFPs was affected by the distribution of charged residues, and found that some of the uptake deficiencies of weakly performing scGFPs can be accounted for using a model that describes the electric field of a protein in solution (**Figure 2.5b**, see Methods for details of model)

Taken together, these findings are consistent with a model in which at least one cellular uptake potency-determining event, such as proteoglycan binding or receptor crosslinking[21], is dependent on the arrangement of cationic groups on the surface of a supercharged protein, and not simply on its net charge.

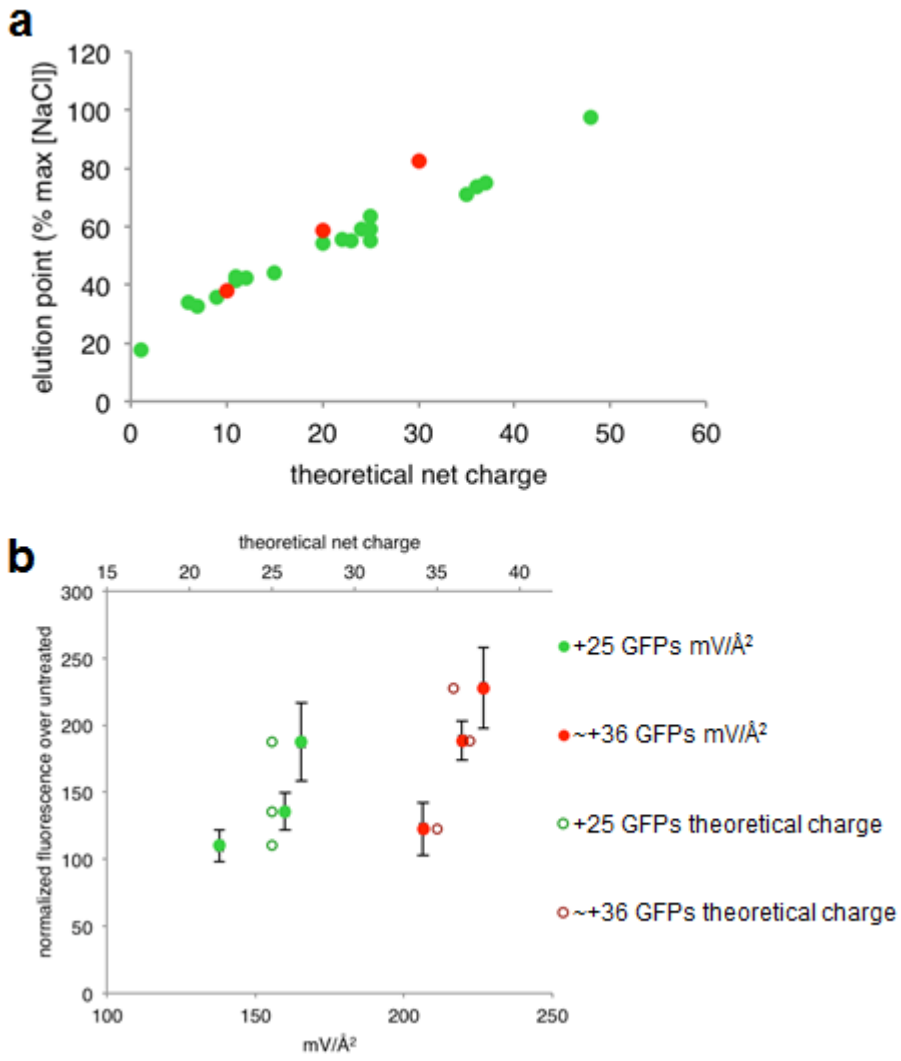

Figure 2.5 The effect of surface charge distribution on actual protein charge and protein voltage. (**a**) Charge-dependent elution of scGFPs and poly-Lys/Arg peptides from cation exchange resin. Each protein was loaded onto a HiTrap SP XL cation exchange column (GE Healthcare) in a 100 μL volume at 50 μM protein concentration. Proteins were eluted over a gradient from 1x PBS to 1x PBS with 1 M NaCl. Protein elution was monitored by absorbance at 280 nm. Green points represent scGFPs. Red points represent poly-Lys/Arg peptides. **(b)** The effect of surface charge distribution on protein voltage and cellular uptake. HeLa cells were treated with 200 nM of one of three +25 GFP variants (+25, +25a, or +25b GFP) or one of three +36 GFP variants (+35, +36, or +37 GFP). A positive relationship between protein voltage and cellular uptake is evident that is not apparent when considering only the theoretical net charge (open data points, top axis).

The cell-penetration properties of macromolecule delivery agents can depend strongly on the associated cargo[12]. To determine whether the observed charge-uptake relationship for scGFPs is sustained for scGFP-mediated delivery of fused proteins, a subset of scGFPs were fused to either mCherry or to Cre recombinase[22,23]. GFPs with theoretical net charges of +9, +15, +24, +25 (three variants designated +25, +25A, and +25B), +35, +36, +37, and +48 were chosen as fusion partners to cover a range of theoretical net charges, and to include some variants with altered charge distributions that performed below the level predicted by their theoretical net charge alone. For comparison, the CPPs Tat, Arg_{10}, and penetratin were also prepared as fusions with mCherry and Cre recombinase.

HeLa cells were incubated with mCherry fusions for 4 hours at 50 nM to 2 μM, washed, and analyzed by flow cytometry using mCherry fluorescence as a measure of protein delivery. The mCherry fusions retained the same general trend of charge-dependent cellular uptake observed for scGFPs alone, with low-potency delivery transitioning to high-potency delivery (up to 50-fold more potent) for variants with > +15 theoretical net charge (**Figure 2.6a**). We also observed lower cell penetration potency of the same high-charge variants that underperformed in the cellular uptake study described above (+25A, +25B, +35, +37, and +48 GFPs). Fusions with CPPs performed as previously observed, with low overall mCherry delivery potency comparable to that of the +9 GFP-mCherry fusion, except for penetratin-mCherry, which resulted in moderate levels of delivery at the highest concentrations tested (1 to 2 μM).

Next we used scGFP-Cre recombinase fusions to assay the functional delivery of protein enzymes to the cytosol. BSR.LNL.tdTomato cells[12] were used to report Cre-dependent recombination. Since scGFP-Cre fusions have little or no recombinase activity until the scGFP moiety is cleaved from the recombinase moiety[12], we first digested the purified proteins with

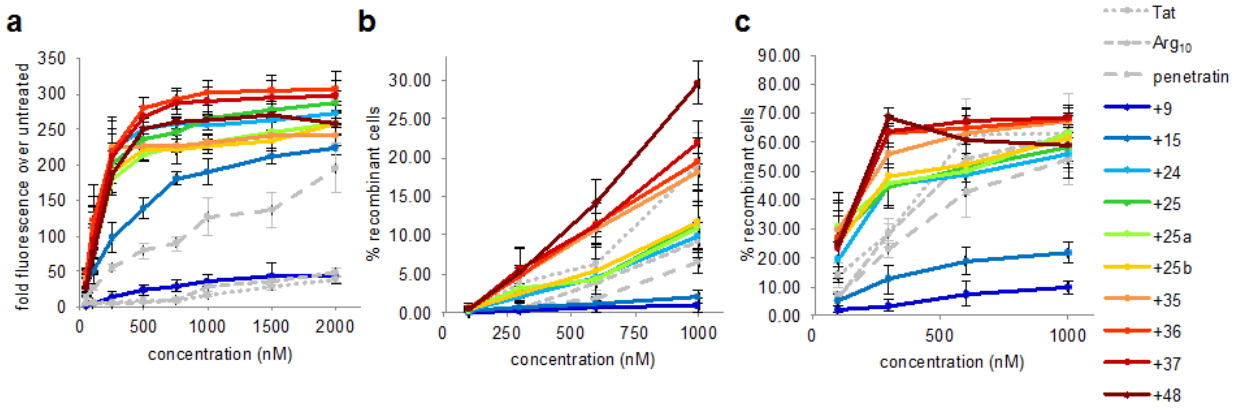

Figure 2.6 Charge-dependence of fused protein uptake and functional delivery. (**a**) HeLa cells were treated with the indicated scGFP-mCherry for 4 hours, washed, and analyzed by flow cytometry. The plot shows median mCherry fluorescence of cells relative to untreated cells. (**b**) BSR.LNL.tdTomato Cre reporter cells were treated with scGFP-Cre fusions for 4 hours incubated an additional 48 hours, and assayed by flow cytometry. (**c**) BSR.LNL.tdTomato reporter cells were treated and analyzed exactly as in (**b**) except 100 μM chloroquine was added during the 4-hour protein incubation and for an additional 12 hours post-treatment.

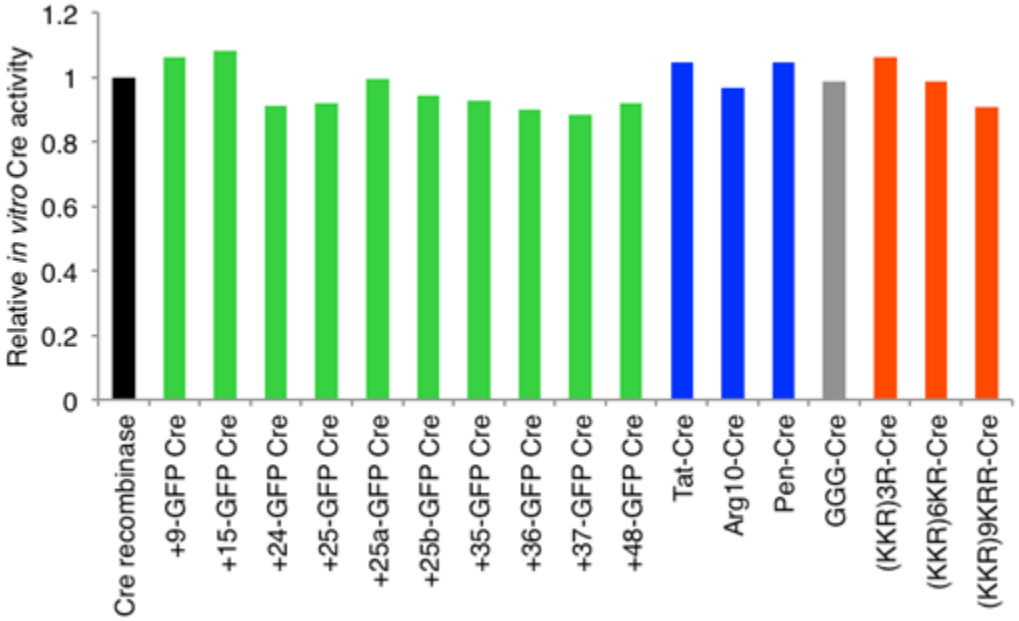

Figure 2.7 *In vitro* activity of Cre protein fusions and peptide conjugates. The activity of purified Cre recombinase fusions and Cre conjugates to poly-Lys/Arg peptides was analyzed using an *in vitro* plasmid recombination assay. Following 1 hour incubation at 37 °C, the reaction was analyzed by gel electrophoresis and the amount of recombined plasmid was determined by densitometry. The data is normalized relative to the activity of wild-type Cre. All proteins were pre-treated with cathepsin B to cleave the fused scGFP domain,

cathepsin B to cleave the linker between scGFP and Cre and assayed the resulting proteins *in vitro* as previously described[12]. All cathepsin B-digested scGFP-Cre fusions exhibited recombinase activities comparable to that of wild-type Cre (**Figure 2.7**). In addition to being cleaved from scGFP, delivered Cre recombinase must also tetramerize, escape endosomes, translocate to the nucleus, and recombine the reporter gene cassette to trigger tdTomato reporter gene expression. Cells were incubated with each of the scGFP-Cre fusions across a range of concentrations for 4 hours, washed to remove non-internalized proteins as before, then incubated further in protein-free media for 48 hours.

The percent of recombinant cells resulting from each treatment was revealed by flow cytometry. The Tat, Arg_{10}, and penetratin Cre fusions exhibited functional recombinase delivery efficiencies in the low-potency regime, consistent with previously reported results[11,12]). Functional Cre delivery by scGFPs, in contrast to mCherry delivery, was strictly charge-dependent, with higher charged scGFPs producing more recombinant cells, and like-charged variants clustering together (**Figure 2.6b**). Interestingly, variants that were observed to be internalized with significantly lower efficiency than +36 GFP, such as +37 GFP and +48 GFP, when fused to Cre resulted in comparable or even greater recombinant cells than +36 GFP, respectively. The differences between the cellular uptake potencies of these scGFPs (**Figure 2.6a**) and their strictly charge-dependent ability to deliver functional Cre (**Figure 2.6b**) suggests that protein-intrinsic factors downstream of internalization, such as cleavage of the fusion or endosomal escape efficiency, are potency-limiting in the latter case.

To test these possibilities, Cre delivery was repeated in the presence of chloroquine, a small molecule that enhances endosomal escape and increases functional delivery potency of

supercharged proteins[12]. Importantly, the profile of charge- and dose-dependent Cre recombination (**Figure 2.6b**) is significantly shifted towards greater functional delivery in the presence of 100 μM chloroquine (**Figure 2.6c**), and closely reflects the pattern of total fusion protein uptake (**Figure 2.6a**).

To probe potential differences in cleavage efficiency of Cre fused to different supercharged proteins, we subjected +15 GFP-Cre and +36 GFP-Cre to *in vitro* and cell-based cleavage assays. Both proteins were found to be equally cleavable (**Figure 2.8a, b**), suggesting that proteolytic cleavage efficiency is not a determinant of the different observed functional delivery abilities. We also assayed the effect of blocking cleavage by co-incubation of BSR reporter cells with +36 GFP-Cre and a broad cathepsin inhibitor, Z-Phe-Gly-NHO-Bz[24]. At low doses (0.1-0.5 μM) there is a clear depression of delivery (**Figure 2.9**), presumably due to a failure to liberate Cre from +36 GFP. At high doses (2-100 μM), the efficiency of Cre delivery is dramatically improved, possibly due to continued inhibition of endosomal proteases that would otherwise degrade Cre.

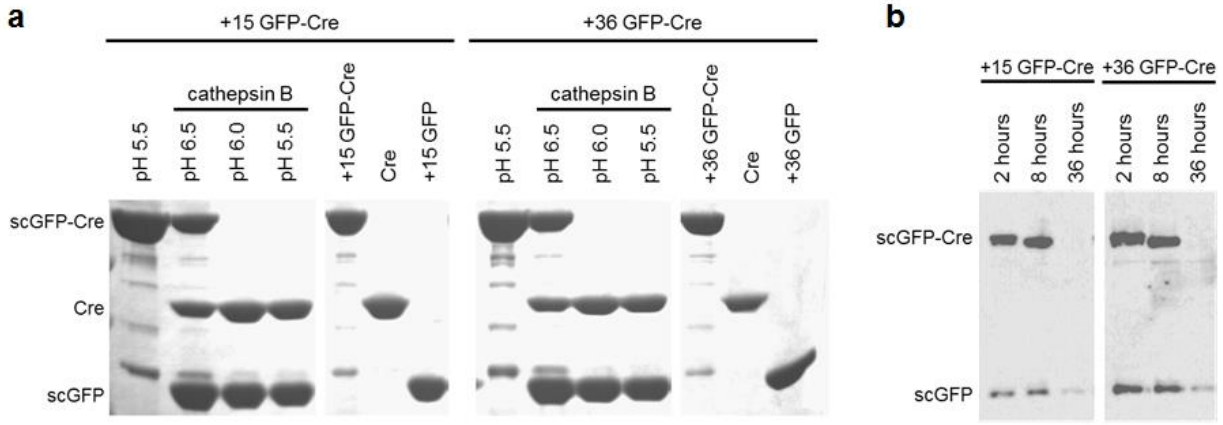

Figure 2.8 Protease susceptibility of scGFP-Cre fusions. (**a**) *In vitro* cathepsin B cleavage assay showing the pH dependence of scGFP-Cre cleavage. Cleavage reactions analyzed by PAGE and stained by Coomassie blue. Pure preparations of scGFP-Cre, Cre, and scGFP proteins are included as references for cleavage products (**b**) Immunoblot of HeLa cell lysates treated for 1 hour with 0.5 μM scGFP-Cre, washed, and then incubated for the indicated period of time prior to harvesting of lysate.

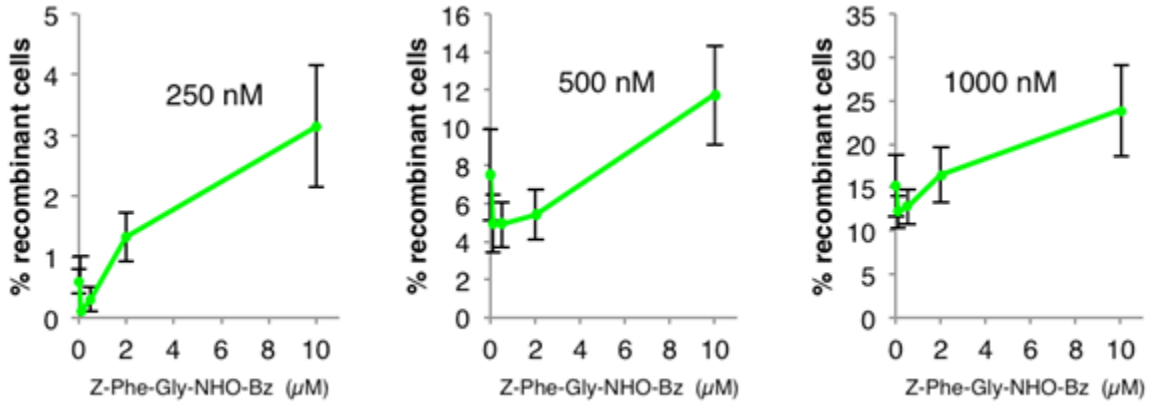

Figure 2.9 The effect of protease inhibition on +36 GFP-Cre delivery. BSR reporter cells were incubated with the indicated dose of +36 GFP-Cre protein and Z-Phe-Gly-NHO-Bz cathepsin inhibitor for 4 hours in serum-free media, and then incubated a further 48 hours prior to analysis of recombinant cells by flow cytometry. Plots show % recombination as a function of inhibitor dose for three +36 GFP-Cre treatments (250 nM, 500 nM, and 1000 nM protein).

Taken together, these results suggest that functional protein delivery by scGFPs is charge-dependent and protease-dependent, and that this dependence is at least partially the result of post-internalization processes. Processes that are likely to influence the intracellular survival and functional delivery of a protein, such as endosomal trafficking, maturation, or escape, may therefore be affected by the presence of highly cationic molecules within endosomes, a hypothesis explored in the experiments below.

The above studies implicate theoretical net charge as a strong determinant of cell penetration and protein delivery potency. Next we sought to determine the effect of protein structure on cell ular uptake and and protein delivery potency by measuring the delivery efficiency of proteins fused to Arg- and Lys-rich peptides with Lys/Arg content comparable to that of scGFPs. Because genes encoding simple fusions of mCherry or Cre to polycationic Arg or Lys peptides longer than 10 amino acids did not express efficiently in *E. coli*, we used an enzyme-mediated protein ligation strategy to generate these proteins. The synthetic peptides

$(KKR)_3R$, $(KKR)_6KR$, and $(KKR)_9KRR$ were efficiently conjugated to mCherry and Cre using a highly active mutant sortase enzyme recently evolved in our laboratory (Supplemental Information Experimental Procedures)[25]. The resulting poly-Lys/Arg-fused mCherry and Cre proteins were incubated with reporter cell lines as described above. For comparison, cells were also treated with mCherry and Cre fused to +9, +15, +25, or +36 GFP, which collectively span the range of theoretical net charges covered by the poly-Lys/Arg-fused proteins. Cellular uptake and protein delivery potencies were quantitated by flow cytometry as before.

For mCherry delivery, scGFP fusions consistently outperformed fusions with similarly charged poly-Lys/Arg-conjugated mCherry (**Figure 2.10a**). This difference was especially pronounced at lower concentrations. While at higher concentrations fusions with +20 and +30 peptides approached the performance levels of fusions with +25 GFP, in most cases and at most concentrations scGFPs resulted in more mCherry delivery than both similarly charged and more highly charged cationic peptides (**Figure 2.10b**). These results indicate that scGFP proteins result in distinctly more potent cellular uptake and mCherry delivery than unstructured cationic peptides, even when the scGFP and the cationic peptides possess a similar theoretical net charge.

The results for Cre recombinase delivery were generally consistent with the mCherry delivery results; scGFP-Cre fusions typically outperformed cationic peptide-Cre fusions, especially at lower doses, with the exception of the lower potency +9 and +15 GFPs (**Figure 2.10c**). Interestingly, the +20 and +30 poly-Lys/Arg peptides exhibited decreased Cre delivery potency compared with +10 poly-Lys/Arg, which is approximately as effective as Arg_{10}-Cre (**Figure 2.6b**), despite the fact that all Cre conjugates and fusions were comparably active *in vitro* following cathepsin B treatment (**Figure 2.7**)[12]. A possible explanation for the lower potency of the +20 and +30 poly-Lys/Arg peptides is their potential cytotoxicity, as has been

previously reported for certain cationic peptides and other synthetic cationic polymers[15,26]. This toxicity may not have manifested on the shorter timescale of the mCherry internalization assays, while the Cre assay is necessarily a multi-day experiment.

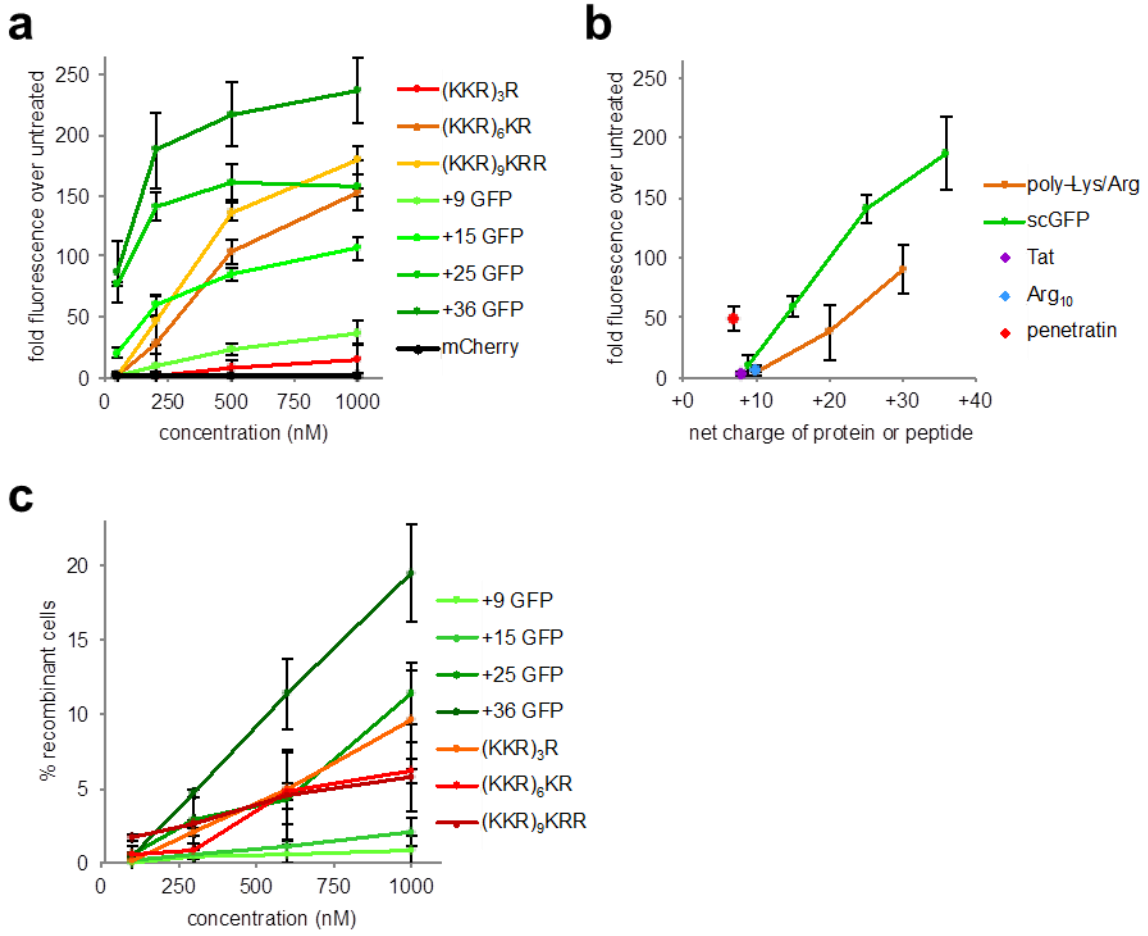

Figure 2.10 scGFPs deliver mCherry and Cre recombinase more effectively than similarly charged cationic peptides. Cells were treated and analyzed as described in **Figure2.6a** and **b**. (**a**) Median mCherry fluorescence of cells treated with scGFP-mCherry fusions or cationic peptide-mCherry conjugates at the indicated doses for 4 hours. (**b**) mCherry delivery efficiency as a function of protein net charge. (**c**) Percent recombinant BSR.LNL.tdTomato reporter cells expressing tdTomato following treatment with scGFP-Cre fusions or cationic peptide-Cre conjugates. Error bars represent the standard deviation of three experiments.

Collectively, these results reveal a cellular uptake and protein-delivery potency advantage from displaying positive charge on a structured protein surface, as is this case with scGFPs, over

simply appending the same number of cationic residues in a simple, presumably unstructured peptide tag. This difference may be explained if the higher density of Arg and Lys side chains in a short peptide prevents the complete protonation of every residue by decreasing their pK_a values to reduce unfavorable charge-charge repulsion, decreasing the actual cationic net charges attainable by synthetic peptides compared with supercharged proteins. Alternatively, spreading cationic residues over a much larger surface area may engage cell-surface receptors or other proteins involved in endocytosis and/or escape from endosomes more effectively than concentrating cationic charge in a small peptide tail. We tested the former possibility using the same cation-exchange-based approach that we used to assess the actual charge magnitudes of scGFPs. Poly-Lys/Arg peptides eluted from the cation exchange resin in a manner consistent with their theoretical net charge (**Figure 2.5a**). These results support a model in which the potency differences between cationic peptides and supercharged proteins arise from differences in their ability to interact with cellular components rather than from differences in their actual versus theoretical net charge.

We previously reported that scGFP uptake is an energy-dependent process that requires actin polymerization and the presence of cell-surface sulfated proteoglycans[13]. Given these findings, one possibility is that scGFP may be taken up via macropinocytosis, similar to some CPPs[27,21]. In light of the observed cellular uptake and macromolecule delivery differences between scGFPs and cationic CPPs, however, we hypothesized that significant mechanistic differences in their cellular internalization may also exist. We therefore probed the uptake of scGFP in greater detail.

We used a collection of known endocytic inhibitors to probe the role of different endocytic pathways in the uptake of scGFP. HeLa cells were pre-treated with each inhibitor for

1 hour prior to incubation with 500 nM +36 GFP and inhibitor for an additional hour (**Figure 2.11a**). Cells were washed as described above and scGFP uptake was measured by flow cytometry.

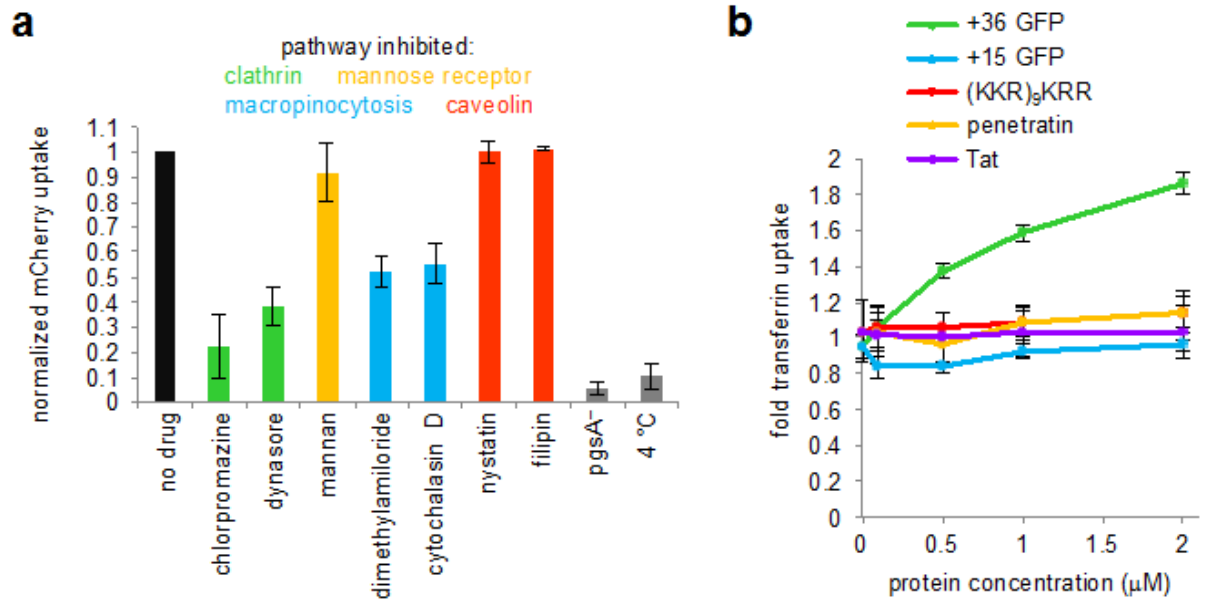

Figure 2.11 Effects of endocytic pathway probes on mCherry delivery by +36 GFP. (**a**) Cells were pre-treated for 1 hour with the indicated inhibitor prior to 1-hour treatment with 500 nM +36 GFP in the continued presence of inhibitor, washed, and analyzed by flow cytometry. A mutant cell line, CHO pgsA^{-}, which lacks heparin sulfate proteoglycans, and treatment at 4 °C instead of 37 °C, were included as controls (grey bars). (**b**) HeLa cells co-treated with Texas Red-labeled transferrin and the indicated cationic protein or peptide for 4 hours were washed and analyzed by flow cytometry.

Amiloride prevents the activation of macropinocytosis[28] that are implicated in the uptake of cationic CPPs, and inhibits Tat peptide delivery nearly completely at 5 mM[27]. In contrast, 5 mM amiloride reduced uptake of scGFP by only 48%. Likewise, cytochalasin D, an inhibitor of actin polymerization, blocked uptake of scGFP by 45% at 10 μM,, but has been reported to block Tat peptide uptake by >90%[27]. These findings suggest that, in contrast to most cationic CPPs, scGFP uptake is not entirely dependent on macropinocytosis and actin-dependent processes. Interestingly, 5 μg/mL chlorpromazine, which prevents the formation of clathrin-coated pits, and 50 μM dynasore, which prevents the scission of clathrin-coated vesicles[29,30], inhibited uptake of

+36 GFP by 78% and 62%, respectively. Finally, nystatin and filipin, two inhibitors of caveolin-dependent uptake which is involved in the internalization of some viral delivery agents and non-viral protein-based agents[31,32], resulted in no observable impact on scGFP uptake at 50 nM of either inhibitor. These inhibitor studies collectively implicate a clathrin-dependent endocytosis as a major mechanism for the uptake of scGFP, and suggest that macropinocytosis plays a lesser role.

To test whether scGFPs modulate clathrin-dependent uptake, we incubated HeLa cells with fluorophore-labeled transferrin, a known clathrin-dependent cargo, and Tat-mCherry, penetratin-mCherry, $(KKR)_9$KRR-mCherry, +15 GFP, or +36 GFP at concentrations from 0.1 to 2 μM (**Figure 2.11b**). Co-incubation with +36 GFP resulted in a marked dose-dependent increase in intracellular transferrin. None of the other proteins tested had an appreciable effect on transferrin uptake. Titration of transferrin (20 to 200 μg/mL) in the presence of 200 nM +36 GFP had little effect on +36 GFP uptake, indicating that +36 GFP does not use the same receptor as transferrin but instead stimulates transferrin accumulation independently. The stimulation of transferrin accumulation by +36 GFP may implicate clathrin in the uptake of scGFP and suggests that high-potency scGFPs potently alter endocytic processes that rely on this pathway. These results also represent a distinction between high-potency scGFPs and other proteins and peptides tested.

Given the partial inhibition of +36 GFP uptake by the macropinocytosis inhibitor amiloride (**Figure 2.11a**), we tested the effect of epidermal growth factor (EGF), a known inducer of macropinocytosis, on +36 GFP uptake. Across a range of concentrations (0.1 to 1 μg/mL), EGF had no significant impact on +36 GFP uptake (**Figure 2.12a**). However, when the internalization of labeled EGF was studied as function of the concentration of +36 GFP, +15

GFP, Tat-mCherry, penetratin-mCherry, and $(KKR)_9KRR$-mCherry (0.1 to 2 μM each), EGF uptake was inhibited by +15 GFP , +36 GFP, and $(KKR)_9KRR$-mCherry by up to 40% (**Figure 2.12b**). This observation indicates that high-potency cell-penetrant proteins such as +36 GFP may either compete with EGF for receptors, strongly induce ligand-independent internalization of EGF receptors through (a known mechanism of EGFR regulation)[33,34], or alter EGFR activation, internalization, or degradation through another mechanism.

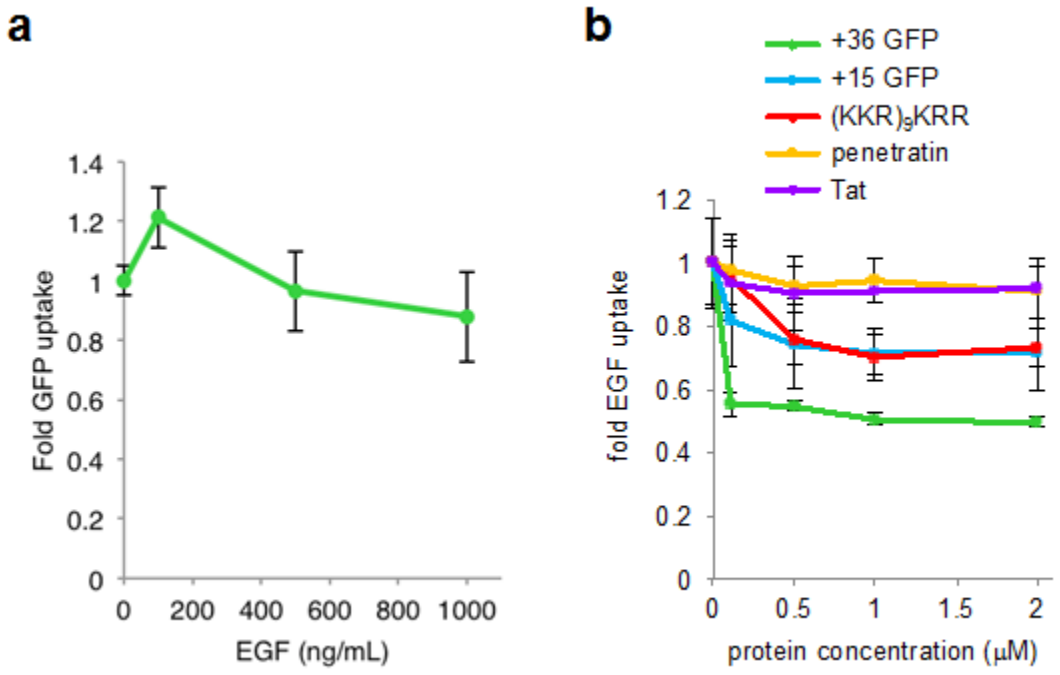

Figure 2.12 Interaction between scGFP and EGF uptake. (**a**) HeLa cells in serum-free DMEM were treated with 200 nM +36 GFP in the presence of EGF ranging from 100 ng to 1 mg/mL, then washed and assayed by flow cytometry to determine GFP uptake. + 36 GFP uptake is normalized to 200 nM +36 GFP uptake in the absence of EGF treatment. (**b**) HeLa cells co-treated with Alexa 594-labeled EGF and the indicated cationic protein or peptide for 4 hours were washed and analyzed by flow cytometry. EGF uptake was normalized to cells treated with Alexa 594-labeled EGF alone. Error bars represent the standard deviation of three experiments.

Collectively, these mechanistic studies indicate that high-potency scGFPs enter cells through uptake pathways that include clathrin-dependent endocytosis, which may be altered by scGFPs, and macropinocytosis. Furthermore, scGFPs cause significant changes to the transport

of both recycling and degradative endocytic cargoes. Such changes correlate with the potency of protein delivery.

The above observations suggest that scGFPs enter cells through specific endocytic routes. Next we characterized the effects of scGFPs on key protein components in endocytosis pathways. Charged peptides including Tat and Arg_{10} have been shown to induce the activation of Rac GTPase[21], an early step required for the induction of macropinocytosis. Rho and Rac are GTPases involved in cytoskeletal reorganization and endosomal trafficking[36,37]. Rac is located near the plasma membrane and initiates actin polymerization at the start of macropinosome formation[38]. Rho is downstream of multiple endocytic pathways, including caveolin-dependent uptake, and clathrin-independent phagocytosis of particulate matter[39]. Rac, on the other hand, is associated with fluid-phase endocytosis[38].

We assayed the activation of the GTP-bound form of Rho and Rac GTPases by scGFPs and poly-Lys/Arg peptides using a colorimetric plate-based "GLISA" (Cytoskeleton, Inc.). The initial binding events of scGFP and cationic peptides depend on association with sulfated proteoglycans[13], and the binding and crosslinking of proteoglycans is known to activate endocytosis[40,41]. We therefore used CHO cells as well as mutant CHO $pgsA^-$ cells defective for xylosyltransferase[42], which is responsible for sulfation of proteoglycans, as a control. Cells were incubated for 5 minutes with proteins, harvested, and lysed. The resulting cell lysates were applied to GLISA plates. Upon treatment with high-potency scGFPs and poly-Lys/Arg peptides, Rho was activated roughly 2-fold compared with untreated controls (**Figure 2.13a**). Low-charge variants resulted in significantly lower activation levels, consistent with the observed sigmoidal charge-uptake relationship for scGFP (**Figure 2.3** and **2.4a**). Treatment of the CHO $pgsA^-$ cells resulted in low Rho activation across all treatments, indicating that Rho activation is dependent

on sulfated proteoglycans (**Figure 2.13a**, dark bars). Importantly, the activation of Rho by scGFPs is consistent with our observation of a +36GFP-dependent intracellular transferrin accumulation (**Figure 2.11b**), as Rho plays an important role in clathrin organization, cargo sorting, and endosome motility[39].

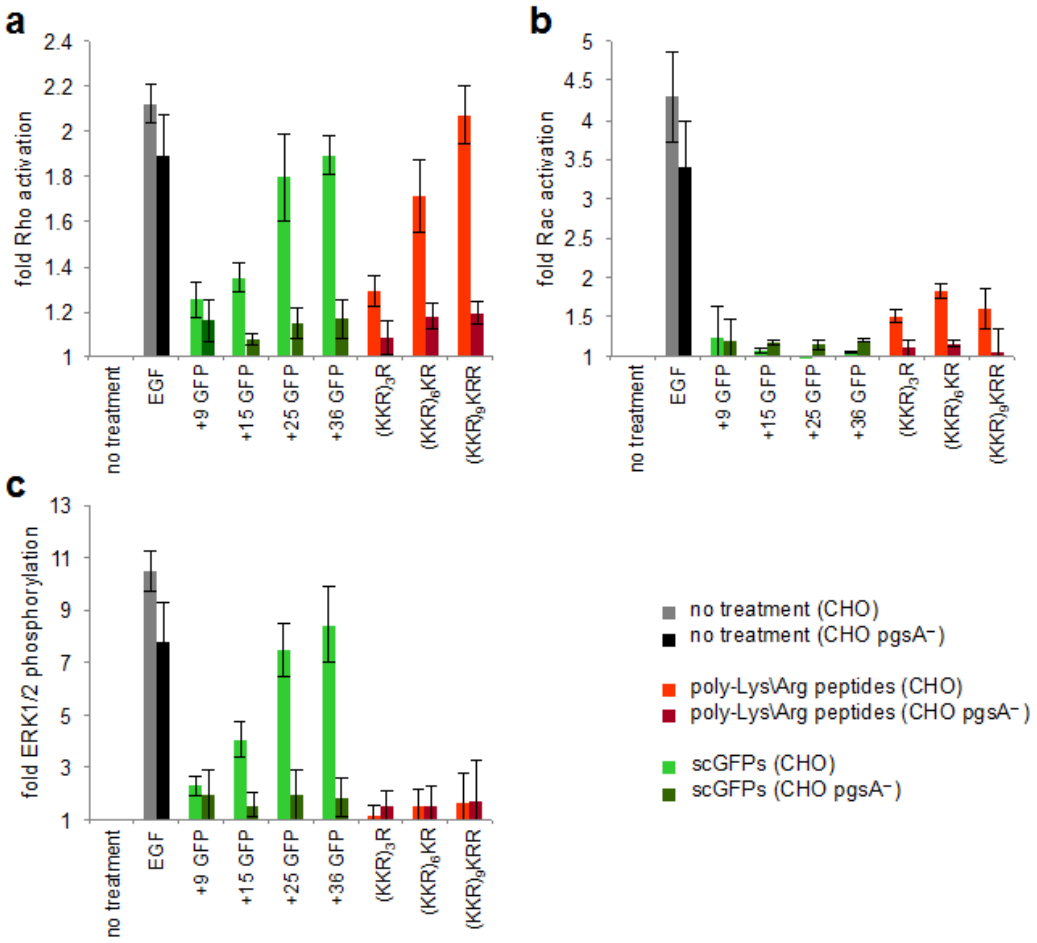

Figure 2.13 activation of Rho-GTP and ERK1/2 phosphorylation by scGFPs and cationic peptides. CHO cells (light bars) or CHO pgsA$^-$ cells (dark bars) were treated with 1 μM of the indicated scGFP or cationic peptide. EGF treatment at 50 ng/mL was included as a control for endocytosis activation ("EGF"). (**a**) Rho-GTP quantitation by GLISA. (**b**) Rac-GTP quantitation by GLISA. (**c**) The extent of ERK1/2 phosphorylation determined by immunoblot with anti-phospho ERK1/2 antibody and quantitated by densitometry. Error bars represent the standard deviation of three experiments.

We also assayed the activation of GTP-bound Rac by GLISA following treatment with scGFPs or poly-Lys/Arg peptides, and observed no activation by scGFPs and weak activation by

poly-Lys/Arg peptides (**Figure 2.13b**), consistent with observations by others that Rac activation increases following treatment with either Arg_8 or Tat peptides[21]. Rho and Rac are known to negatively regulate one another[43], and significant Rho activation may depress Rac levels. Given the role of Rho in the activation and maturation of different endocytic pathways[39,38], the above results could provide a molecular explanation for the scGFP-mediated alteration of transferrin and EGF endocytosis (**Figure 2.11b** and **2.12b**).

In many cell lines, endocytosis is required for full activation of ERK1/2 following receptor ligation[44,45,46]. Therefore, changes in ERK1/2 phosphorylation can reflect activation of plasma membrane signaling receptors by scGFPs or changes in the endocytic transport of these receptors[47].

We measured ERK1/2 phosphorylation by immunoblot analysis of cell lysates treated for 5 minutes with scGFP or poly-Lys/Arg variants to quantitate phospho-ERK1/2 levels. Upon treatment with highly charged scGFPs, but not similarly charged poly-Lys/Arg peptides, cellular ERK1/2 was phosphorylated by up to 8.4-fold over controls lacking treatment (**Figure 2.13c**). More modestly charged (+9 or +15) scGFPs also induced ERK1/2 phosphorylation, although to a lower extent (**Figure 2.13c**). In all cases tested, ERK 1/2 phosporylation was also dependent on the presence of sulfated proteoglycans, as evidenced by the lack of ERK1/2 phosphorylation in the CHO $pgsA^-$ cells (**Figure 2.13c**, dark bars). This finding suggests that the activation of receptor signaling, and possibly receptor internalization, is an important factor in scGFP uptake. These results also provide additional mechanistic distinctions of scGFP uptake compared with the uptake of unstructured cationic peptides, which have no significant effect on ERK1/2 phosphorylation.

The amount of material per endosome, the rate of intracellular transport and maturation, and the ultimate destination of endocytosed material are all parameters known to vary among endocytic routes[48]. These distinctions can impact the characteristics and effectiveness of protein delivery. A comparison of scGFP uptake, scGFP-mCherry uptake, and functional Cre delivery (**Figures 2.3, 2.4a,** and **2.6a-c**) suggests charge-dependent differences following internalization that contribute to the effectiveness of extraendosomal functional delivery of protein cargoes. The high charge of cationic delivery reagents could alter the ionic composition of the endosomal lumen. As the trafficking of endosomes is closely tied to the function of endosomal membrane ion channels and pumps[49], significant alteration through such a mechanism could disrupt endosome transport and maturation. Endocytosed cargoes targeted for degradation are transported to perinuclear lysosomal compartments within approximately 2 hours[50]. The effective cytosolic delivery of endocytosed material requires that it avoids degradation in the endocytic pathway. A difference in the transport to lysosomes between scGFPs and cationic peptides may explain the observed variations in protein delivery data described above.

We monitored the transport of proteins to lysosomes following internalization by incubation of HeLa cells with 500 nM of a protein of interest for 1 hour. Following this incubation, cells were washed extensively with PBS containing heparin, and incubated a further 4 hours in protein-free media containing heparin to prevent the uptake of any remaining extracellular protein and inhibit the reinternalization of material from recycling endosomes. Dextran, a marker of fluid-phase endocytosis[51], was included as a control. Lysosomes were labeled with the LysoTracker reagent (Life Technologies), which effectively labels acidified lysosomes.

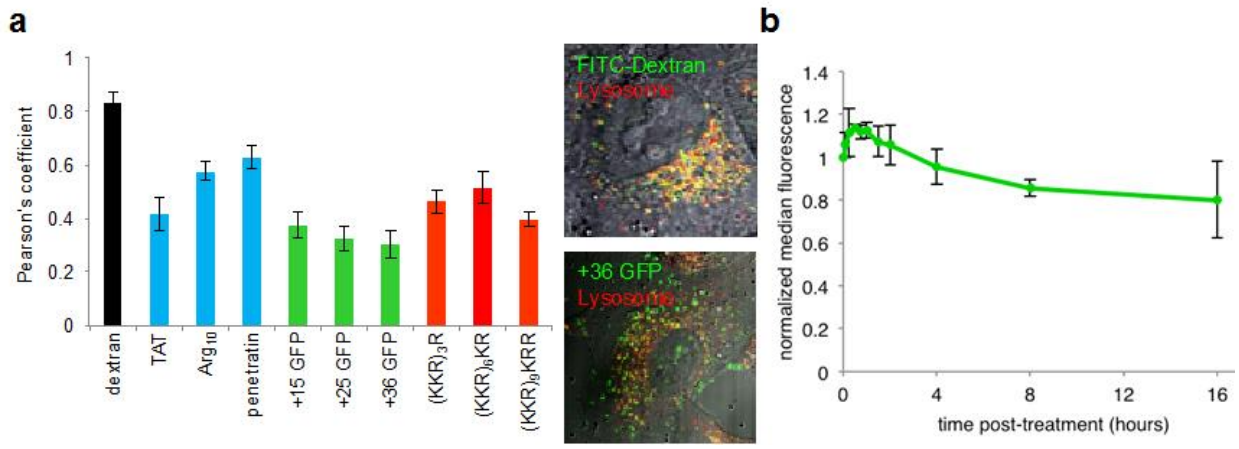

Figure 2.14 Lysosomal localization of scGFPs and cationic peptides. (**a**) HeLa cells were treated with scGFP proteins or peptide-mCherry fusions for 1 hour, washed, incubated an additional 4 hours and imaged by confocal microscopy. Lysosomes were labeled with LysoTracker Red or Green as appropriate (Life Technologies). The colocalization of proteins with lysosomes was determined by calculation of the Pearson's correlation coefficient of the red and green channels using ImageJ (left). A representative image of cells treated with either FITC-dextran (top right) or +36 GFP (bottom right) is shown. (**b**) Long term survival of scGFP within endosomes. HeLa cells treated with +36 GFP for 1 hour were washed and incubated for up to 16 additional hours in DMEM with 1% FBS to prevent dilution through cell division and 20 U/mL heparin to prevent continued endocytosis of residual extracellular +36 GFP protein. Cells were analyzed by flow cytometry, and normalized by the median fluorescence of cells immediately following the 1-hour incubation (time = 0). Error bars represent the standard deviation of three experiments.

We performed co-localization analysis of Lysotracker dye with the scGFPs and cationic peptide-tagged mCherry proteins to determine the extent of endosomal trafficking or maturation disruption (**Figure 2.14a**). Dextran efficiently co-localized with lysosomes, resulting in a Pearson's correlation coefficient (ρ) of 0.8. In contrast, all delivered proteins displayed some level of significantly slower transport, with none exhibiting $\rho > 0.6$. In all cases, large numbers of protein-containing peripheral endosomes were observed, even 4 hours after removal of the protein reagent from the media. Such a failure to localize to perinuclear acidic vesicles suggests either a slowing of the maturation of protein-containing endosomes or the sorting of scGFPs away from the degradative pathway. The observation of a peripheral vesicles was not due to continued or recent internalization of proteins since the cells were incubated in protein-free media following washing and prior to imaging. Likewise, the apparent peripheral distribution

was not due to the degradation and disappearance of otherwise perinuclear lysosomal material over time, as the total cellular fluorescence did not significantly decrease over the course of the 4-hour incubation following treatment (**Figure 2.14b**).

The extent of lysosomal colocalization was significantly lower for all scGFPs compared with any cationic peptide tested (**Figure 2.14a**). The scGFPs exhibited the least lysosomal co-localization with a trend of decreased lyosomal localization with increased charge. Within the CPP set of Tat, Arg_{10}, and penetratin, the trend of colocalization strongly reflected functional delivery properties, with Tat (the most potent delivery agent of the cationic peptides in this study) exhibiting the lowest lysosomal colocalization. The colocalization data for the poly-Lys/Arg peptides, in contrast, did not follow this trend. While all poly-Lys/Arg peptides showed significantly reduced lysosomal co-localization compared to dextran, they were among the poorest performing domains in the Cre recombinase delivery assays (**Figure 2.10**).

Collectively, these lysosomal localization results reveal that the relative delivery abilities of scGFPs and CPPs correlate with the extent of slowed endosomal maturation, and that scGFP exhibits a stronger effect on these processes than CPPs. That such an effect correlates with the potency of functional delivery suggests that the ultimate non-endosomal fate of delivered proteins relies at least in part on how they are trafficked after internalization.

After observing the strong effect on endosome transport to lysosomes among the cationic delivery reagents tested, we characterized the events of early endocytosis and tracked differences between scGFPs and CPPs leading up to the observed late-stage endosomal distribution using a recently reported high-throughput confocal microscopy image analysis pipeline(**Figure 2.15**, see Methods for detailed description of analysis pipeline)[48]. This platform enables the measurement of a wide range of endosomal parameters, including trafficking kinetics, endosome number, size

distribution, and content. As representative proteins, we applied this strategy to the study of +36 GFP (a high-potency supercharged protein), +15 GFP (a modestly potent supercharged protein), Tat-mCherry, penetratin-mCherry, and +30 $(KKR)_9KRR$-mCherry (three cationic peptides of varying charge and size)

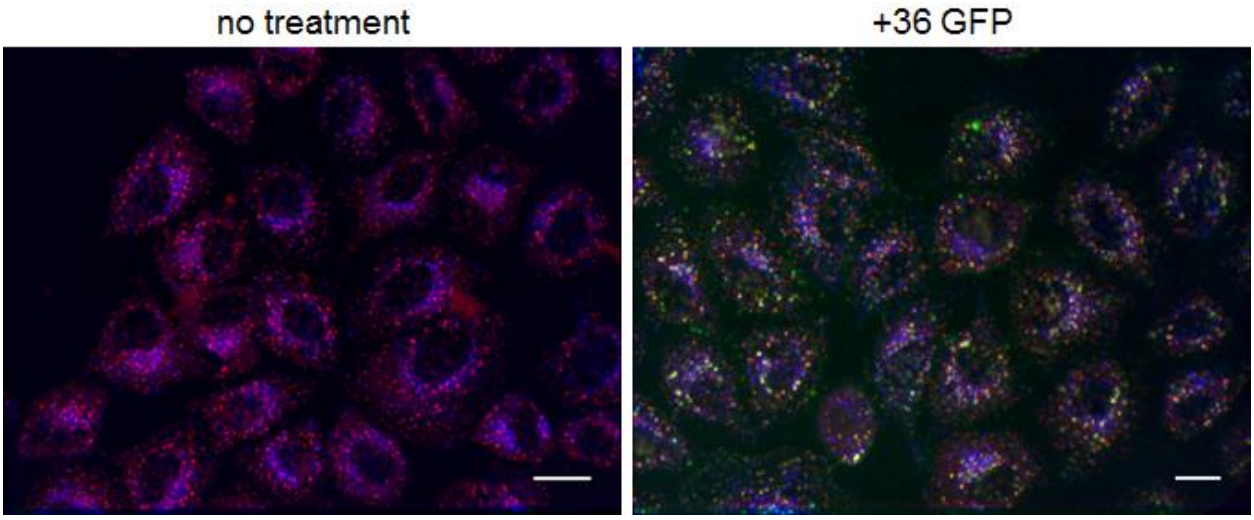

Figure 2.15 High throughput confocal microscopy images subjected to trafficking analysis. Sample of representative images assayed for localization of scGFPs with early or late endosomal markers. HeLa cells treated with 500 nM of the indicated protein and subjected to high throughput confocal microscopy as EEA1 is pseudo-colored red, LAMP1 is pseudo-colored blue, and +36 GFP signals are pseudo-colored green. Scale bars represent 10 micrometers.

HeLa cells were incubated with 0.1 to 2 μM of each protein for 30 minutes, washed, and fixed. Over a range of concentrations, the number of scGFP or CPP-containing vesicles increased steadily (**Figure 2.16a**). Among the proteins tested, +36 GFP occupied the highest number of endosomes, even at low concentration (0.1 μM), while other proteins did not occupy such a high number of endosomes even at the highest dose (2 μM). The +30 poly-Lys/Arg peptide exhibited a decrease in the number of endosomes formed at high doses. Tat was only detected in a small number of endosomes, consistent with its modest cell-uptake potency.

Importantly, over a range of protein concentrations, the amount of protein per endosome did not exhibit a significant dose-dependent increase (**Figure 2.16b**). Endosomes with +36 GFP

contained roughly 4- to 10-fold more protein per endosome than endosomes with CPPs, while +15 GFP-containing endosomes had roughly 2-fold more protein than CPPs. These findings have important implications for interpreting the concentration dependence of functional protein delivery. While the amount of protein per endosome is likely a factor contributing to the overall effectiveness of protein functional delivery, as evidenced by the relative amounts of protein per endosome for +36 GFP and +15 GFP versus CPPs (**Figure 2.16b**), the number of endosomes containing endocytosed protein, rather than the amount of protein per endosome, more closely correlates with dose-dependent changes in functional delivery potency. Increasing the dose of protein during treatments resulted in both more protein-containing endosomes and more potent functional delivery (**Figure 2.16a** and **2.6b**, respectively).

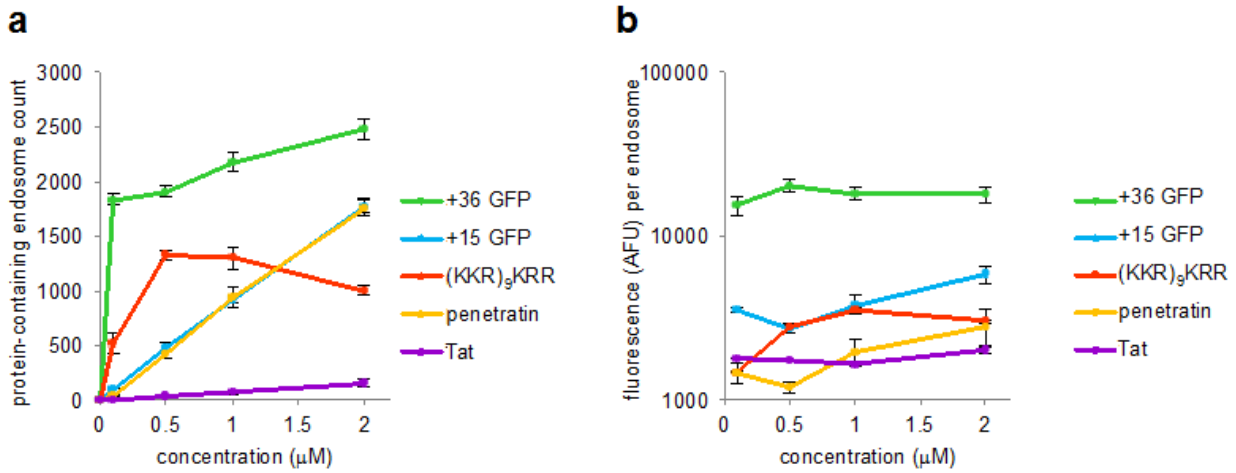

Figure 2.16 Effect of scGFPs and cationic peptide concentration on endosome formation and endosomal protein content. HeLa cells were incubated with the indicated concentrations of scGFP or peptide-mCherry fusion for 1 hour, washed, and the number of GFP or mCherry-containing endosomes was analyzed by high throughput microscopy (**a**) Supercharged proteins and cationic peptides are endocytosed with greatly different potencies. (**b**) High-potency scGFPs fill individual endosomes with more protein than less potent scGFPs and cationic peptides.

We next measured the uptake and trafficking of proteins over time. Cells were incubated with 500 nM proteins for up to 2 hours. +36 GFP rapidly entered cells, filling a maximal number of early endosomes (identified by the presence of EEA1[52]) within 10 minutes of treatment (**Figure 2.17**). The other proteins were much slower in their uptake, requiring as much

as 2 hours before reaching maximal early endosome occupancy. Thereafter, +36 GFP quickly exited early endosomal compartments. The rapid uptake and transport to early endosomes of +36 GFP in comparison with the other proteins may reflect the ability of this protein to be mainly internalized by clathrin-mediated endocytosis, as described above (**Figure 2.14a**).

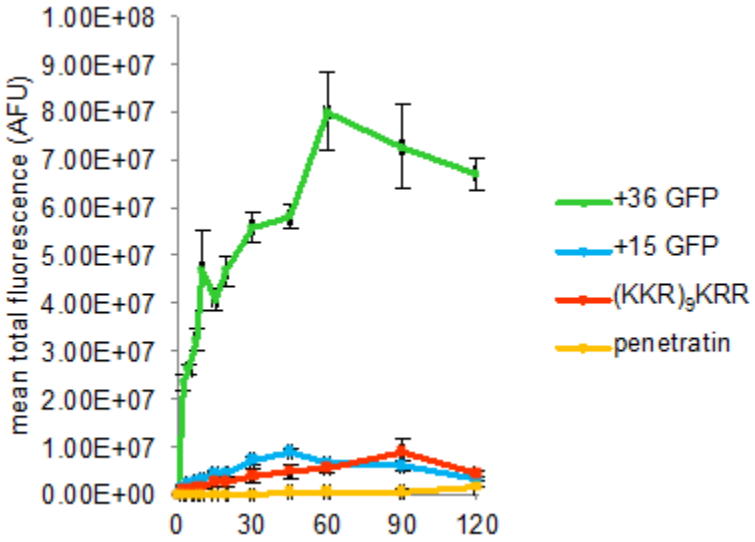

Figure 2.17 Uptake and kinetics scGFPs and cationic peptides. HeLa cells were incubated with the indicated scGFP or peptide-mCherry fusion at 500 nM for up to 2 hours and analyzed by confocal microscopy.

Over the course of two hours, a small fraction (< 20%) of +36 GFP eventually localized to late endosomes compartments (identified by the presence of LAMP1) (**Figure 2.18**). Interestingly, the vast majority of +36 GFP protein is unaccounted for in either of these major endosomal populations. The other proteins tested were slow to enter both early endosomal compartments and late endosomal compartments, while a major portion of the protein was

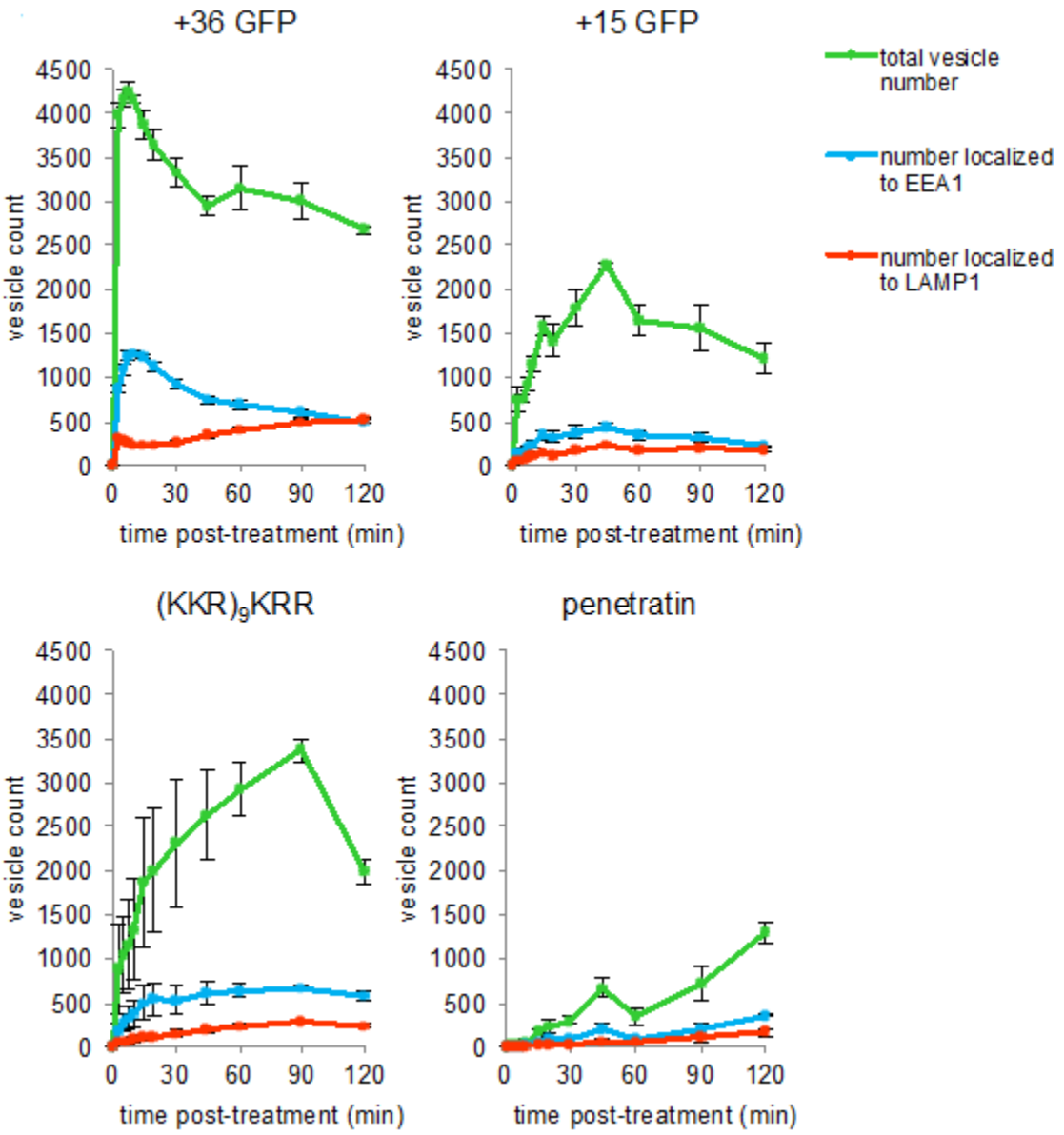

Figure 2.18 Trafficking kinetics of scGFPs and cationic peptides. Trafficking of protein through early endosomes (labeled with anti-EEA1 antibody) and lysosomes (labeled with anti-LAMP1 antibody). The number of GFP- or mCherry-containing vesicles showing EEA1 or LAMP1 colocalization was an analyzed by high-throughput automated confocal microscopy. In all plots, green points represent the total number of vesicles containing scGFP or peptide-mCherry fusions; blue points are vesicles showing EEA1 colocalization; and red points are vesicles showing LAMP1 colocalization. All error bars represent the standard deviation of three experiments.

retained outside of either of these endosomal populations, as with +36 GFP. The amount of protein degradation post-endocytosis, as measured by loss of fluorescence, was not appreciable for any of the proteins within the 2 hours tested. The lack of degradation over this period, and

even over a longer 16-hour incubation (**Figure 2.14b**), suggests a role for dramatically slower lysosomal degradation in functional protein delivery. Indeed, 80% of the original +36 GFP signal remained within cells after 16 hours, indicating very little proteolysis of the endocytosed cargo. In contrast, the intracellular lifetime of other endosomal cargo is comparatively short. For example, transferrin can be recycled out of the cell rapidly, with an intracellular half-life of 7 minutes[53]; EGF is almost completely degraded within 2 hours[54]; and polystyrene beads are transported to lysosomes within 2 hours[50]. The presence of intact protein within an endosomal reservoir for several hours post-internalization may provide a much greater opportunity for these proteins to escape endosomes, even through a low-efficiency mechanism, than endosomal cargoes that are rapidly degraded through canonical pathways.

Finally, we measured the amount of protein per endosome and the size of endosomes over time. These two parameters are indicative of endosomal maturation and fusion, and may determine the long-term fate of endocytosed proteins. Over 2 hours the average amount of protein in each cell did not decrease for any protein tested. However, unique among all proteins tested, the number of +36 GFP-containing endosomes decreased sharply (**Figure 2.18**). Consequently, the amount of +36 GFP per endosome increased dramatically over time (**Figure 2.19a**). This protein concentration effect was much less pronounced for the other proteins tested, with +36 GFP reaching > 5-fold more protein per endosome than any other agent tested, and achieving a 5-fold increase in the amount of protein per endosome over 2 hours. The other agents exhibited a much more modest increase in the amount of protein per endosome during this time (**Figure 2.19a**). These measurements suggest that +36 GFP is not significantly inhibiting homotypic fusion and cargo accumulation in endosomes. The size of protein-containing

endosomes did not change over time appreciably for any protein tested (**Figure 2.19b**), and all endosomes were smaller compared to both early and late endosomes.

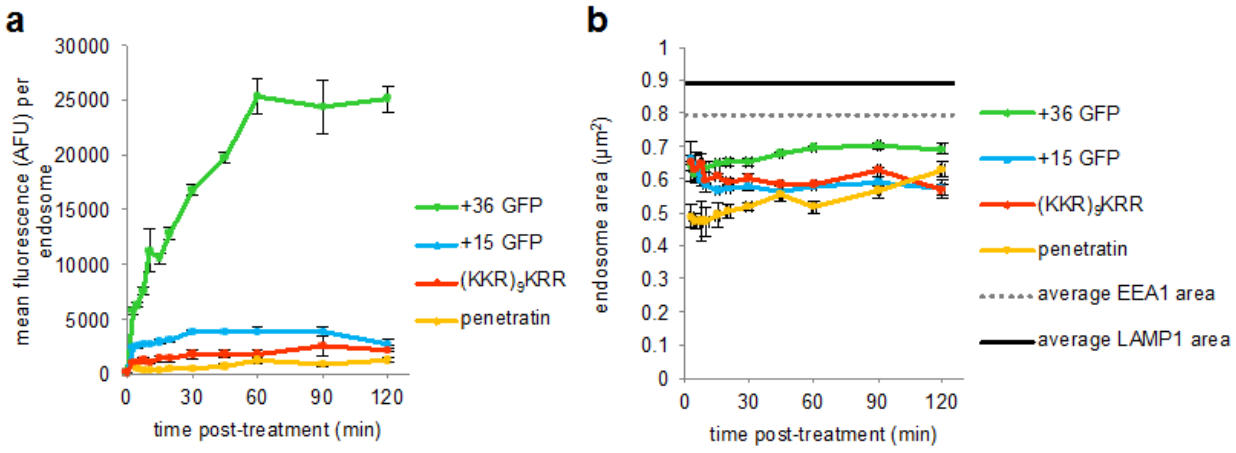

Figure 2.19 Time-dependence of endosomal protein content and size following treatment with scGFPs and cationic peptides. (**a**) Amount of GFP or mCherry protein per endosome as a function of time. (**b**) Size of endosomes as a function of time, with the average size of early endosomes and lysosomes indicated by the dotted and solid lines, respectively. All error bars represent the standard deviation of three experiments.

These observations suggest a role for the long-term intracellular survival of proteins within a unique peripheral endosomal population during the process of functional protein delivery by supercharged proteins and cationic peptides. The endosomes occupied by +36 GFP are indeed acidified exit the early unacidified endosomal population, as marked by EEA1, yet they are not effectively transported to late endosomes (**Figure 2.18**) or lysosomes (**Figure 2.14a**). Our observations for +36 GFP may reflect a unique trafficking alteration that is shared at least in part by other polycationic delivery reagents. The specific cause and consequence of these changes is not obvious, but it is possible that a high concentration of cationic moieties within an endosome disrupts lumenal ion composition, preventing the maturation of protein-containing endosomes and their subsequent fusion with degradative compartments. If the intra-

endosomal survival of proteins is extended as a result, then the delivery of functional extra-endosomal proteins may strongly depend on such alterations to endosomal trafficking.

Finally, we have developed a method to adapt existing liposomal nucleic acid delivery techniques to the delivery of proteins.

The surface chemistry of proteins varies dramatically between distinct protein sequences, making general, one-size-fits-all delivery formulations difficult to achieve. While superpositively charged proteins can drive efficient internalization due to their extreme charge, the nature of the fused cargo protein can sometimes dominate the chemical properties and performance of the fusion protein. Issues with solubility in physiologic conditions, sensitivity to degradation, and the potential immunogenicity of naked recombinant proteins may limit their broad application.

In contrast nucleic acids can be treated as comparatively identical in terms of their bulk chemical properties due to their more uniform polymeric structure that is held in common even between highly divergent sequences. Nucleic acid delivery, when compared to protein delivery technologies is a much more mature technology. The widespread, routine use of commercial cationic polymers and lipids is a testament to this. Liposomes in particular are an attractive delivery platform, as the nature of lipid bilayer structure makes membrane-membrane fusion within endocytic vessels a natural and efficient means of endosomal escape. Liposomes further stabilize and protect their cargo until the point of delivery, potentially reducing immunogenicity, degradation, and the effective dose of cargo delivered.

We hypothesized that the endosomal escape properties inherent to liposomal formulations could be brought to bear on protein delivery if a method to drive efficient encapsulation of

proteins could be developed. Cationic lipid reagents drive efficient encapsulation of nucleic acids due to electrostatic interactions with the anionic phosphate backbone. To mimic this interaction, we tested whether -30 GFP, originally developed as a counterpart to +36 GFP in the study of protein folding, could drive encapsulation of a fused cargo protein and mediate functional delivery via liposomes.

We constructed a translation fusion of -30 GFP to Cre recombinase and assayed delivery as above using the BSR.LNL.tdTomato reporter cell line. As an initial trial, Lipofectamine RNAiMAX (Life Technologies), an siRNA delivery reagent, was used to attempt encapsulation of the purified recombinant protein. This particular liposomal formulation was chosen due to the comparable charge and size of siRNAs to -30 GFP, rather than plasmid DNA delivery reagents which are optimized for larger cargoes and greater negative charge. Encapsulation protocols were followed exactly as recommended for siRNA cargo, substituting the purified protein instead.

In intial tests, -30 GFP-Cre was found to be efficiently delivered by the liposomal formulation across a wide range of doses (**Figure 2.20**). Surprisingly, the optimal dose of -30 GFP-Cre when used with liposomes was 50-fold lower than that of naked +36 GFP Cre protein. Indeed, optimal delivery of -30 GFP Cre induced recombination in 50% of cells, compared to only 15% from +36 GFP Cre at a 50-fold higher treatment. Interestingly, liposomal encapsulation drove more effient delivery of both +36 GFP-Cre and nonfused Cre recombinase, consistent with the general ability of liposomes to enhance endosomal escape. However, only -30 GFP-Cre was delivered efficiently at low doses. At doses as low as 5 nM, -30 GFP-Cre is able to induce recombination at a rate comparable to that of a 1 μM +36 GFP-Cre treatment, a 200-fold improvement in delivery.

These remarkable results have prompted us to explore the delivery of other protein cargoes, including potential therapeutics, and to further optimize both the encapsulation process and liposomal formulation for eventual application *in vivo*.

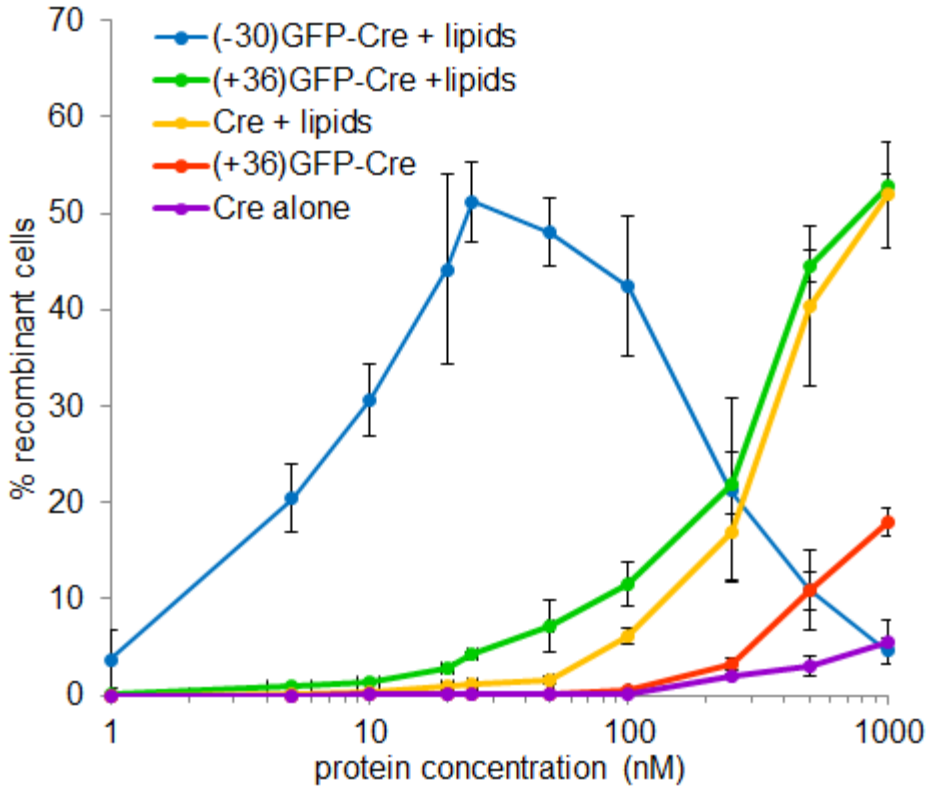

Figure 2.20 -30 GFP fusion enables efficient liposomal delivery of proteins. The indicated protein was incubated with 1.5 uL of LipoFectamine RNAiMAX across a range of concentrations and subsequently used as per the manufacturers protocol. Cells were treated for 4 hours, media was replaced, and cells were subsequently analyzed by flow cytometry to detect recombinants.

Discussion

In this study we found that the cellular uptake ability of supercharged GFPs exhibits a strong sigmoidal charge dependence. Such a relationship suggests distinct interactions with cellular components or uptake through different endocytic routes by high-charge, high-potency scGFPs compared with CPPs or low-potency scGFPs. Indeed, our subsequent mechanistic

studies revealed that high-potency scGFPs alone require clathrin and dynamin for efficient uptake, and modify the transport of clathrin-dependent endocytic cargoes. The potency of scGFPs correlates with the level of activation of Rho and ERK1/2. Such activation may involve the crosslinking of sulfated proteoglycans or other receptors on the cell surface, a process known to induce macropinocytosis by CPPs[55]. If receptor crosslinking is indeed the basis of endocytic activation by supercharged proteins, then perhaps the extended surface area provided by highly charged variants explains their greater observed potency compared with more modestly charged scGFPs or unstructured, short cationic CPPs.

The endocytosis inhibitor and transferrin uptake studies described above suggest that scGFP cellular uptake relies significantly on the induction of clathrin-dependent uptake. While internalization of scGFPs is still dependent on the presence of anionic sulfated proteoglycans, the studies above suggest that the major route of scGFP uptake is not macropinocytosis. These results, combined with the mechanistic studies implicating Rho and ERK1/2 activation, suggest that scGFP uptake likely proceeds through multiple pathways and depends most heavily on a proteoglycan-requiring, clathrin-dependent process.

We observed that both scGFPs and CPPs undergo an altered process of trafficking following endocytosis. The delivery agents tested here are not effectively transported to degradative lysosomal compartments. Moreover, the endocytosed proteins are localized within abnormally small peripheral endosomes that are negative for the early endosome marker EEA1. The magnitude of the changes to endosomal transport correlates with the ability of each reagent to deliver cytosolic proteins. This novel finding has important implications for the survival of delivered proteins within cells, and for macromolecular delivery vehicles in general. If protein-containing endosomes fail to mature and efficiently fuse with lysosomal compartments, then

delivered proteins may be provided with an extended temporal window during which they can escape into the cytoplasm. Such a phenomenon could contribute to the effectiveness of all cationic delivery vehicles, especially high-potency supercharged proteins, which altered endosome trafficking more than the other proteins and peptides studied here. The exact mechanism by which scGFPs and CPPs disrupt endosomal transport is not yet understood and represents an attractive subject of future studies. Supercharged GFPs also caused significant changes to the transport of the endocytic cargoes transferrin and EGF, raising the possibility that supercharged proteins have multiple effects on specific molecular components of the endocytic system that lead to changes in transferrin receptor and EGF receptor recycling and degradation.

Together, our findings reveal new insights into the mechanism of cellular uptake and trafficking of supercharged proteins and CPPs, and more generally highlight the importance of studying at the subcellular level the interaction of delivery agents with their target cells. Such studies improve our understanding of how macromolecule delivery agents work, and identify means by which to optimize their function. These studies also suggest ways in which exposure to such delivery agents can alter cell physiology, an important consideration as macromolecular medicines become increasingly pervasive human therapeutics.

Methods

Construction of scGFPs

We used a simple PCR and restriction enzyme-based approach to rapidly generate each scGFP variant with a common cloning strategy. Plasmids encoding stGFP, +15 GFP, +25 GFP, and +36 GFP, and +48 GFP were used as templates for PCR to generate three fragments spanning the full-length GFP sequence coding sequence. The three classes of PCR products were bounded by pairs of endogenous restriction enzyme cleavage sites (*Nhe*I/*Nde*I, *Nde*I/*Eag*I, or *Eag*I/*Xho*I). The appropriate sets of PCR

products were digested with the requisite enzyme pair and ligated into a pET vector backbone to generate pET-+*N* GFP, where *N* is the theoretical net charge of the resultant protein calculated by summing the charge of all acidic and basic residues. The choice of restriction sites did not alter the sequence of the resulting protein.

Construction of scGFP-mCherry and scGFP-Cre fusions

A pET expression vector was prepared containing a $(GGS)_9$ linker flanked at the 5' by *Nhe*I and *Age*I sites, and *Kpn*I and *Bam*HI at the 3'. The coding sequences for mCherry and Cre were PCR amplified and cloned into the *Kpn*I and *Bam*HI site, to place them C-terminal of the linker. The appropriate scGFP sequence was then cloned N-terminal of the linker into *Nhe*I and *Age*I sites to generate the sets of scGFP-mCherry and scGFP-Cre expression constructs.

Generation of poly-Lys/Arg-linked mCherry and Cre using an evolved sortase enzyme

For proteins with N-terminal peptides, pET expression plasmids were generated by PCR amplification of the coding sequence of either mCherry or Cre using primers that add a triglycine sequence to the N-terminus (MGGG-, where M should be efficiently removed by *E. coli* processing machinery), and a 6xHis tag sequence to the C-terminus.

For proteins used with C-terminal peptide tags, pET expression plasmids were generated by PCR amplification of the coding sequence of either mCherry or Cre, with primers to add the sequence –LPETGGSHHHHHH to the C-terminus. Such a sequence is a substrate for sortase, which removes the 6xHis tag, and appends GGG-poly-Lys/Arg peptide to generate –LPETGGG-poly-Lys/Arg.

Poly-Lys/Arg peptides were synthesized by the Tufts University Proteomics Core. For N-terminal peptides, the poly-Lys/Arg sequence was immediately followed by the sequence LPETGG, for sortase recognition, cleavage and ligation to GGG-terminated proteins. For C-terminal peptides, the poly-Lys/Arg sequence was preceded by GGG to enable it to act as the nucleophile in the sortase tagging reaction, resulting in ligation to the C-terminus of the protein.

All sortase tagging reactions took place using the following conditions: 100 μM of the appropriate N-terminal GGG- or C-terminal LPETG-bearing protein, 2 mM of the appropriate LPETG- or GGG-

bearing peptide, 5 mM $CaCl_2$, 5 μM evolved sortase enzyme, 1xTBS, incubated at 25 °C for 1 hour. Following the incubation, EDTA was added to 1 mM and reactions were placed on ice for 5 minutes. Unconjugated peptide was depleted by concentrating and re-diluting the reactions three times in 30 kDa cutoff microcentrifuge concentrator tubes (Millipore), for a total of ~1,000-fold dilution of the excess peptide. The reaction was then purified by cation exchange to remove the unreacted proteins or peptides and the sortase enzyme.

Protein expression and purification

Briefly, *E. coli* BL21(DE3) was transformed with the appropriate pET expression vector. Purification involved His-tag affinity chromatography using Ni-NTA agarose, followed by cation exchange to remove contaminating anionic *E. coli* components that may confound downstream mammalian cell-based assays. Protein purity was monitored by SDS-PAGE and Coomassie Blue staining.

Protein uptake and delivery assays

Briefly, scGFP and scGFP-mCherry delivery assays were performed with HeLa cells. Proteins were diluted in serum-free DMEM and incubated on the cells in 48-well plates for 4 hours at 37 °C. Following incubation, the cells were washed three times with PBS + 20 U/mL heparin and trypsinized for analysis of protein uptake by flow cytometry.

Cre delivery assays used the BSR.LNL.tdTomato cell line. Cre fusion proteins were diluted in serum-free media on the cells in 48-well plates for 4 hours at 37 °C. Following treatment, the cells were incubated a further 48 hours in serum-containing media prior to trypsinization and analysis by flow cytometry.

All flow cytometry were carried out on a BD Fortessa flow cytometer (Becton-Dickinson) using 530/30 nm and 610/20 nm filter sets.

Rho, Rac, and ERK1/2 activation

Rho and Rac activation assays were performed using colorimetric Rho and Rac GLISA kits (Cytoskeleton). CHO and CHO-pgsA$^-$ cells were plated in 60 mm plates at 4 x 10^4 cells per plate. Cells

were grown for 24 hours in DMEM-F12 with 1% FBS, and then 16 hours in serum-free DMEM-F12 to serum-starve prior to treatment. Media was aspirated from cells, and the appropriate protein treatment diluted in pre-warmed serum-free DMEM-F12 was applied to the cells in plates. Immediately following a 5-minute incubation, cells were placed on ice, media was aspirated, and cells were washed three times in ice cold PBS + heparin to remove surface bound proteins. Cells were lysed in 100 μL of the provided lysis buffer, scraped from the plate, transferred to chilled 1.5 mL tubes and immediate frozen in liquid nitrogen. The assays were performed exactly as recommended by the manufacturer, using 30 μL of lysate per well of the GLISA plate.

For phosphorylation of ERK1/2, lysates prepared as described above were diluted with 4xLDS sample buffer (Invitrogen), boiled for 5 minutes, and 12 μL loaded into 12% Bis-Tris NuPAGE gels (Invitrogen). Following electrophoresis on 12% NuPage Bis-Tris gels (Invitrogen), Western blot using a mouse anti-pERK1/2 antibody (Cell Signaling) at a dilution of, and a rabbit anti-beta-tubulin antibody (AbCam) as loading control. Secondary detection was done using a goat anti-rabbit 800 and goat anti-mouse 680LT antibodies (LiCor). Blots were analyzed using a Odyssey Imager (LiCor)

Subcellular localization assays

HeLa cells were plated in 60 mm glass slide-bottomed plates at 4 x 10^4 cells per plate. 24 hours later, cells were treated for 1 hour with 500 nM of the appropriate protein diluted in serum-free DMEM. Following incubation, cells were washed three times with PBS + 20 U/mL heparin to remove extracellular protein. Cells were incubated a further 4 hours in serum-containing DMEM + 20 U/mL heparin to allow endocytosed proteins to reach a trafficking endpoint. Cells were treated with either LysoTracker RED or LysoTracker GREEN (Invitrogen), at a concentration of 75 nM for the final 30 minutes of the 4 hour incubation to label lysosomes. Following incubation, media was aspirated and cells were washed once with PBS + 20 U/mL heparin, before replating fresh serum-containing media + 20 U/mL heparin.

Cells were imaged at the HCBI (Harvard University) on an LSM 510 confocal microscope (Zeiss) using 530/20 and 620/40 filters. 15 individual cells per treatment were photographed. TIFF images were

analyzed with ImageJ using the WCIF-ImageJ Collection (Wright Cell Imaging Center) using default parameters.

Surface charge density analysis

After stripping the crystal structure of superfolder GFP (PDB Code 2B3P) of all water molecules and non-protein atomic coordinates, hydrogen atoms were assigned to the structure by the Visual Molecular Dynamics protein visualization package and then minimized in 100 steps of steepest descent structure minimization in the NAMD molecular dynamics package using generalized born electrostatics and the CHARMM force field, with only the installed hydrogens allowed to vary. This optimized structure was then used for all subsequent analyses.

Peptide structures were initially generated using the Molefacture tool in VMD, and then equilibrated by a 1 ns molecular dynamics simulation using the NAMD molecular dynamics package at 310 K using generalized born implicit solvation and the CHARMM force field. A snapshot of the protein structure was then taken every 100 ps of a 9 ns molecular dynamics simulation and structurally aligned to generate a structural ensemble of each peptide. These structures were then overlaid and their voltage expectation values calculated as described below, with a single modification: after calculation of the voltage expectation value $\langle V \rangle$ and square-voltage expectation value $\langle V^2 \rangle$, each value was normalized to the number of structures in the ensemble to generate approximations of $\langle V \rangle$ and $\langle V^2 \rangle$ which take into account the intrinsic flexibility of these structures.

To determine the overall charge density of a supercharged protein sequence, the protein was first reduced to a point cloud representation. From this the convex hull, Γ of the sfGFP protein structure was determined using the TetGen tetrahedral mesh generator package in Mathematica, as was the protein center of mass, ρ. The expanded convex hull Γ′ was then generated by radially expanding each point of Γ by 5 Å from ρ ($\Gamma' = \left\{\forall \gamma \epsilon \Gamma : \overrightarrow{\gamma'} = \vec{\rho} + \frac{\vec{\gamma}-\vec{\rho}}{\|\vec{\gamma}-\vec{\rho}\|}(\|\vec{\gamma}-\vec{\rho}\|+5)\right\}$). Each backbone atom was then assigned a charge according to the canonical partial charges of backbone atoms in a peptide chain (N-**H** 0.2, **N**-H -0.2, **C**=O 0.42, C=**O** -0.42). Charged residues were assumed to have charge isotropically distributed

around their Cα atoms, and so each residue was treated as a point charge centered at their respective Cα coordinate. The electrical potential function V was then given by:

$$V(\vec{x}) = \frac{1}{4\pi\varepsilon\varepsilon_0}\sum_{i=1}^{N}\frac{q_i}{\|\vec{r_i} - \vec{x}\|}$$

Where $\vec{r_i}$ is the atomic coordinate of the i-th charged atom, q_i is its charge, ε is the permittivity of the medium in question (here, ε=81 for liquid water) and ε_0 is the permittivity of the vacuum. The expectation value for voltage $\langle V \rangle$ and its square $\langle V^2 \rangle$ over Γ' were then computed by the integrals

$$\langle V \rangle = \frac{\int_{\Gamma'} V(x) d\Gamma'}{\int_{\Gamma'} d\Gamma'}$$

$$\langle V^2 \rangle = \frac{\int_{\Gamma'} V^2(x) d\Gamma'}{\int_{\Gamma'} d\Gamma'}$$

And the standard deviation of this expectation value was then computed using the relation

$$\sigma_V^2 = \langle V^2 \rangle - \langle V \rangle^2$$

Treatment of cells for high-throughput confocal microscopy

96-well μ clear black flat-bottom polystyrene plates were purchased from Greiner-bio-one; Transferrin-Alexa 647 and EGF-Alexa 555 were purchased from Invitrogen. Rabbit polyclonal anti-EEA1 was previously described; mouse monoclonal anti-LAMP1 was purchased from BD Pharmingen. All fluorochrome-labeled secondary antibodies were purchased from Invitrogen. CO_2-independent medium was purchased from Invitrogen.

HeLa cells grown in 96-well plates were incubated at different time intervals with 500 nM of +36GFP, +15 GFP, TAT-mCherry, Pen-mCherry or $(KKR)_9$KRR-Cherry at 37 °C in CO_2-independent medium for up to two hours prior to washing with 20 U/mL heparin in PBS and fixation with 4% PFA. Afterwards, cells were permeabilized with saponin (PBS, 0.1% saponin, 4% fish cold gelatin) and stained with antibodies against EEA1 and LAMP1. The antigens were detected with the appropriate fluorophore–labeled secondary antibodies. Cells were stained with DAPI to identify nuclei. All treatments were performed in triplicate.

HeLa cells grown in 96-well plates were stimulated with either 200 ng/mL EGF-Alexa 555, or 25 μg/mL Transferrin-Alexa 647, together with different concentrations of +36 GFP, +15 GFP, TAT-mCherry, penetratin-mCherry, or $(KKR)_9KRR$-mCherry (0, 0.1, 0.5, 1, or 2 μM) for 30 minutes at 37 °C in CO_2-independent medium prior to washing with 20 U/mL heparin in PBS and fixation with 4% PFA. Cells were stained with DAPI to identify nuclei. All treatments were performed in triplicate.

High-throughput confocal microscopy image acquisition and analysis

Images were acquired with an automated spinning-disk confocal microscope (OPERA, Evotec Technologies-PerkinElmer) with a 40X / 0.9 NA water immersion objective. GFP and Transferrin-647 were excited simultaneously with 488 and 635 nm lasers and detected with 520/35 nm and 700/90 nm filters, respectively. mCherry or EGF-Alexa 555 and DAPI were excited simultaneously with 561 and 405 nm lasers and detected with 605/40 nm and 450/50 nm filters. One hundred images per time point were acquired. Every image contained on average 20 cells. To minimize laser variability, all plates were acquired on the same imaging session.

Image analysis was performed using MotionTracking, a custom image analysis software designed by Dr. Kalaidzidis. Before analysis, images were corrected for uneven field illumination and chromatic misalignments using reference images. For every image, individual vesicles in each channel were identified by fitting a sum of powered Lorenzian functions whose coefficients describe the features (e.g. intracellular position, size, fluorescent integral intensity) of each individual objects. The number of vesicles and the total integral intensity were normalized by the cytoplasmic area per frame, to account for frame-to-frame variability in the number of cells. Object-based colocalization was assessed on the basis of cross-sectional overlap between two or three channels. An object scored as colocalized when the total overlap was greater than 40% of the main object cross-section.

To compare the fluorescence intensity of GFP and mCherry in each plate, a solution of 200 nM of +36GFP and 200 nM mCherry was acquired using the same imaging settings. The intensity ratio between mCherry and GFP measured from this solution was used to normalize all GFP intensity values.

References

1. Leader, B., Baca, Q. J. & Golan, D. E. Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov* **7,** 21–39 (2008).

2. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **5,** 993–996 (2006).

3. Schaffert, D. & Wagner, E. Gene therapy progress and prospects: synthetic polymer-based systems. *Gene Ther* **15,** 1131–1138 (2008).

4. Gu, Z., Biswas, A., Zhao, M. & Tang, Y. Tailoring nanocarriers for intracellular protein delivery. *Chem Soc Rev* **40,** 3638–3655 (2011).

5. Wadia, J. S. & Dowdy, S. F. Modulation of cellular function by TAT mediated transduction of full length proteins. *Curr. Protein Pept. Sci.* **4,** 97–104 (2003).

6. Heitz, F., Morris, M. C. & Divita, G. Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. *Br. J. Pharmacol.* **157,** 195–206 (2009).

7. Song, E. *et al.* Antibody mediated in vivo delivery of small interfering RNAs via cell-surface receptors. *Nat. Biotechnol.* **23,** 709–717 (2005).

8. Rizk, S. S. *et al.* An engineered substance P variant for receptor-mediated delivery of synthetic antibodies into tumor cells. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 11011–11015 (2009).

9. Hasadsri, L., Kreuter, J., Hattori, H., Iwasaki, T. & George, J. M. Functional protein delivery into neurons using polymeric nanoparticles. *J. Biol. Chem.* **284,** 6972–6981 (2009).

10. Voelkel, C. *et al.* Protein transduction from retroviral Gag precursors. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 7805–7810 (2010).

11. Cronican, J. J. *et al.* A class of human proteins that deliver functional proteins into mammalian cells in vitro and in vivo. *Chem. Biol.* **18,** 833–838 (2011).

12. Cronican, J. J. *et al.* Potent delivery of functional proteins into Mammalian cells in vitro and in vivo using a supercharged protein. *ACS Chem. Biol.* **5,** 747–752 (2010).

13. McNaughton, B. R., Cronican, J. J., Thompson, D. B. & Liu, D. R. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 6111–6116 (2009).

14. Lawrence, M. S., Phillips, K. J. & Liu, D. R. Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **129,** 10110–10112 (2007).

15. Mitchell, D. J., Kim, D. T., Steinman, L., Fathman, C. G. & Rothbard, J. B. Polyarginine enters cells more efficiently than other polycationic homopolymers. *J. Pept. Res.* **56,** 318–325 (2000).

16. Daniels, D. S. & Schepartz, A. Intrinsically cell-permeable miniature proteins based on a minimal cationic PPII motif. *J. Am. Chem. Soc.* **129,** 14578–14579 (2007).

17. Lee, S.-J., Yoon, S.-H. & Doh, K.-O. Enhancement of gene delivery using novel homodimeric tat peptide formed by disulfide bond. *J. Microbiol. Biotechnol.* **21,** 802–807 (2011).

18. Rothbard, J. B. *et al.* Arginine-rich molecular transporters for drug delivery: role of backbone spacing in cellular uptake. *J. Med. Chem.* **45,** 3612–3618 (2002).

19. Turcotte, R. F., Lavis, L. D. & Raines, R. T. Onconase cytotoxicity relies on the distribution of its positive charge. *FEBS J.* **276,** 3846–3857 (2009).

20. Fuchs, S. M. & Raines, R. T. Arginine grafting to endow cell permeability. *ACS Chem. Biol.* **2,** 167–170 (2007).

21. Nakase, I. *et al.* Interaction of Arginine-Rich Peptides with Membrane-Associated Proteoglycans Is Crucial for Induction of Actin Organization and Macropinocytosis†. *Biochemistry* **46,** 492–501 (2006).

22. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat. Biotechnol.* **22,** 1567–1572 (2004).

23. Guo, F., Gopaul, D. N. & van Duyne, G. D. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **389,** 40–46 (1997).

24. Demuth, H.-U. *et al.* N-peptidyl, O-acyl hydroxamates: comparison of the selective inhibition of serine and cysteine proteinases. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1295,** 179–186 (1996).

25. Chen, I., Dorr, B. M. & Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 11399–11404 (2011).

26. Godbey, W. T., Wu, K. K. & Mikos, A. G. Size matters: molecular weight affects the efficiency of poly(ethylenimine) as a gene delivery vehicle. *J. Biomed. Mater. Res.* **45,** 268–275 (1999).

27. Wadia, J. S., Stan, R. V. & Dowdy, S. F. Transducible TAT-HA fusogenic peptide enhances escape of TAT-fusion proteins after lipid raft macropinocytosis. *Nat. Med.* **10,** 310–315 (2004).

28. Dangoria, N. S., Breau, W. C., Anderson, H. A., Cishek, D. M. & Norkin, L. C. Extracellular simian virus 40 induces an ERK/MAP kinase-independent signalling pathway that activates primary response genes and promotes virus entry. *J. Gen. Virol.* **77 (Pt 9),** 2173–2182 (1996).

29. Wang, L. H., Rothberg, K. G. & Anderson, R. G. Mis-assembly of clathrin lattices on endosomes reveals a regulatory switch for coated pit formation. *J. Cell Biol.* **123,** 1107–1117 (1993).

30. Macia, E. *et al.* Dynasore, a cell-permeable inhibitor of dynamin. *Dev. Cell* **10,** 839–850 (2006).

31. Rothberg, K. G. *et al.* Caveolin, a protein component of caveolae membrane coats. *Cell* **68,** 673–682 (1992).

32. Rothberg, K. G., Ying, Y. S., Kamen, B. A. & Anderson, R. G. Cholesterol controls the clustering of the glycophospholipid-anchored membrane receptor for 5-methyltetrahydrofolate. *J. Cell Biol.* **111,** 2931–2938 (1990).

33. West, M. A., Bretscher, M. S. & Watts, C. Distinct endocytotic pathways in epidermal growth factor-stimulated human carcinoma A431 cells. *J. Cell Biol.* **109,** 2731–2739 (1989).

34. Goh, L. K., Huang, F., Kim, W., Gygi, S. & Sorkin, A. Multiple mechanisms collectively regulate clathrin-mediated endocytosis of the epidermal growth factor receptor. *J. Cell Biol.* **189,** 871–883 (2010).

35. Duchardt, F., Fotin-Mleczek, M., Schwarz, H., Fischer, R. & Brock, R. A comprehensive model for the cellular uptake of cationic cell-penetrating peptides. *Traffic* **8,** 848–866 (2007).

36. Lua, B. L. & Low, B. C. Activation of EGF receptor endocytosis and ERK1/2 signaling by BPGAP1 requires direct interaction with EEN/endophilin II and a functional RhoGAP domain. *Journal of Cell Science* **118,** 2707 –2721 (2005).

37. Qualmann, B. & Mellor, H. Regulation of endocytic traffic by Rho GTPases. *Biochem J* **371,** 233–241 (2003).

38. West, M. A., Prescott, A. R., Eskelinen, E. L., Ridley, A. J. & Watts, C. Rac is required for constitutive macropinocytosis by dendritic cells but does not control its downregulation. *Curr. Biol.* **10,** 839–848 (2000).

39. Ellis, S. & Mellor, H. Regulation of endocytic traffic by Rho family GTPases. *Trends in Cell Biology* **10,** 85–88 (2000).

40. Wittrup, A. *et al.* ScFv Antibody-induced Translocation of Cell-surface Heparan Sulfate Proteoglycan to Endocytic Vesicles. *Journal of Biological Chemistry* **284,** 32959 –32967 (2009).

41. Dehio, C. *et al.* Ligation of Cell Surface Heparan Sulfate Proteoglycans by Antibody-Coated Beads Stimulates Phagocytic Uptake into Epithelial Cells: A Model for Cellular Invasion byNeisseria gonorrhoeae. *Experimental Cell Research* **242,** 528–539 (1998).

42. Esko, J. D., Stewart, T. E. & Taylor, W. H. Animal cell mutants defective in glycosaminoglycan biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **82,** 3197–3201 (1985).

43. Sander, E. E., ten Klooster, J. P., van Delft, S., van der Kammen, R. A. & Collard, J. G. Rac Downregulates Rho Activity. *J Cell Biol* **147,** 1009–1022 (1999).

44. Pierce, K. L., Maudsley, S., Daaka, Y., Luttrell, L. M. & Lefkowitz, R. J. Role of endocytosis in the activation of the extracellular signal-regulated kinase cascade by sequestering and nonsequestering G protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 1489–1494 (2000).

45. Robertson, S. E. *et al.* Extracellular signal-regulated kinase regulates clathrin-independent endosomal trafficking. *Mol. Biol. Cell* **17,** 645–657 (2006).

46. Ung, C. Y. *et al.* Simulation of the regulation of EGFR endocytosis and EGFR-ERK signaling by endophilin-mediated RhoA-EGFR crosstalk. *FEBS Lett.* **582,** 2283–2290 (2008).

47. Sorkin, A. & von Zastrow, M. Endocytosis and signalling: intertwining molecular networks. *Nat Rev Mol Cell Biol* **10,** 609–622 (2009).

48. Collinet, C. *et al.* Systems survey of endocytosis by multiparametric image analysis. *Nature* **464,** 243–249 (2010).

49. Huotari, J. & Helenius, A. Endosome maturation. *EMBO J* **30,** 3481–3500 (2011).

50. Blanchette, C. D. *et al.* Decoupling Internalization, Acidification and Phagosomal-Endosomal/lysosomal Fusion during Phagocytosis of InlA Coated Beads in Epithelial Cells. *PLoS ONE* **4,** e6056 (2009).

51. Oliver, J. M., Berlin, R. D. & Davis, B. H. Use of horseradish peroxidase and fluorescent dextrans to study fluid pinocytosis in leukocytes. *Meth. Enzymol.* **108,** 336–347 (1984).

52. Rink, J., Ghigo, E., Kalaidzidis, Y. & Zerial, M. Rab conversion as a mechanism of progression from early to late endosomes. *Cell* **122,** 735–749 (2005).

53. Ghosh, R. N., Gelman, D. L. & Maxfield, F. R. Quantification of low density lipoprotein and transferrin endocytic sorting HEp2 cells using confocal microscopy. *J. Cell. Sci.* **107 (Pt 8),** 2177–2189 (1994).

54. Carpenter, G. & Cohen, S. 125I-labeled human epidermal growth factor. Binding, internalization, and degradation in human fibroblasts. *J. Cell Biol.* **71,** 159–171 (1976).

55. Imamura, J. *et al.* Single Particle Tracking Confirms That Multivalent Tat Protein Transduction Domain-induced Heparan Sulfate Proteoglycan Cross-linkage Activates Rac1 for Internalization. *J. Biol. Chem.* **286,** 10581–10592 (2011).

56. Smith, B. A. *et al.* Minimally cationic cell-permeable miniature proteins via alpha-helical arginine display. *J. Am. Chem. Soc.* **130,** 2948–2949 (2008).

Chapter 3:

Development of a Phage-Assisted Continuous Evolution selection for Site-Specific Recombinases

Abstract

The use of nuclease-based technologies that enable site-directed knockout of genes in intact, living cells, has expanded dramatically in recent years. However, technologies that enable highly efficient gene repair and targeted integration are comparably underdeveloped. Site-specific recombinase enzymes are an attractive basis for a hypothetical gene replacement and and genomic integration technology. Yet there is currently no recombinase platform that is both highly active and readily reprogrammable. To address the limitations of site-specific recombinase programmability, we have developed a genetic selection based on the Phage-Assisted Continuous Evolution (PACE) system, recently developed in our lab, to enable the directed evolution of recombinase proteins towards user-definable target sequences. We have established a deletion-based selection, optimized the performance of the system, and validated its ability to select library members with desirably altered activities. Using this system we have evolved Cre recombinase to higher activity on its native loxP substrate, and have redirected a population of Cre variants towards an asymmetric target site located within the human ROSA26 "safe harbor" locus.

Introduction

The use of genetic modification technologies has increased rapidly in recent years following the discovery and development of protein classes with programmable DNA-binding specificities. Engineered Zinc Finger libraries[1], TALE arrays[2], and most recently the short RNA-guided CRISPR-derived Cas9 system[3], have enabled increasingly facile means of targeting arbitrary genomic sequences. Genetic knockout technologies based on the retargeting of nucleases have proven to be a highly effective and general approach, and have even seen

application in clinical settings.[4] Nuclease-based approaches to genome modification rely on endogenous cellular repair processes to mediate the generation of the desired sequence change.[5,6] In most cases, inactivation of a target sequence is all that is required, and nucleases serve very well in this capacity.[7] Following cleavage of sites within the genome, double strand DNA break repair responses are triggered, and through repeated cleavage, error-prone repair pathways, such as non-homologous end joining (NHEJ),[8] can introduce random insertions or deletions at the target site, frequently inactivating the gene of interest.

Targeted integration and gene repair can also be triggered using this approach, as cells can utilize homologous pieces of DNA provided *in trans* through a process known as homology directed repair (HDR)[6,9,10], to generate a recombinant, user-defined sequence change. However, HDR rates are generally low, and while the efficiency of integration may currently serve as a powerful research tool, higher and more reliable integration rates may be needed for application of gene correction in therapeutic settings.

HDR rates vary greatly between cell types, throughout the course of the cell cycle, and across the genome[5,6]. Indeed, high rates of homologous recombination are only achievable in cells artificially stalled during specific high-HDR stages of the cell cycle, or in abnormal cell types that exhibit high constitutive rates of homologous recombination[6,11,12,10]. Most cells in the body are post-mitotic, and exist in a state that has intrinsically low rates of HDR[6,13]. While methods to augment or stimulate HDR within cells may be developed, it is not clear what impact this may have on the physiological state of target cell[14], or what unintended rare recombination or translocation events it may induce.

One potential way to circumvent the stochasticity of HDR, and the potential danger of augmenting it globally, is the use of site-specific recombinases. Recombinases directly catalyze strand exchange and ligation between DNA molecules, offering an approach to efficient genomic integration. Due to the catalytic mechanism of many recombinases, including the widely used tyrosine recombinases Cre and Flp, induce less DNA damage and toxicity than similarly expressed nucleases. These enzymes are covalently linked the to the DNA backbone via a phospotyrosine during catalysis, and hold the freed 3'OH ends entirely within the recombinase tetrameric complex (**Figure 3.1**), inaccessible to cellular DNA damage sensors.[15] While high level and long term expression of recombinases has been been found to have undesireable effects on genome stability,[16] if used transiently as is the practice with site-directed nucleases, such toxicity may be avoided.

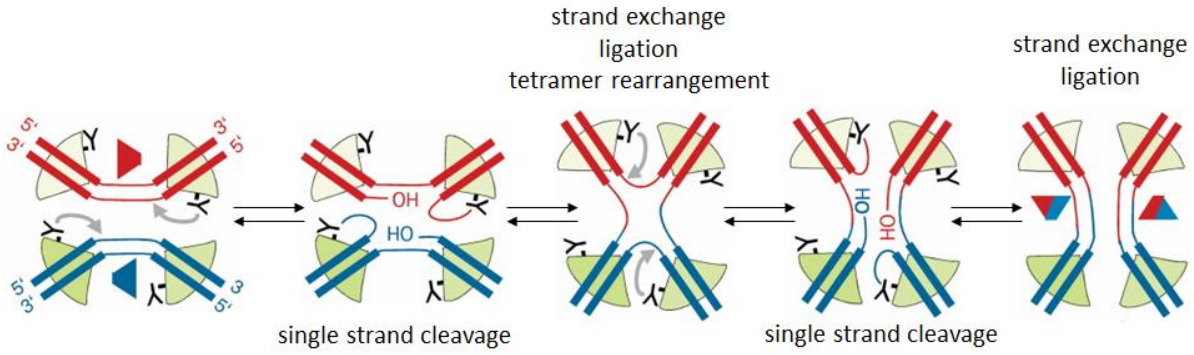

Figure 3.1 Mechanism of recombinase catalyzed strand exchange. The enzyme must assemble as a tetramer before any strand cleavage can occur, so off-target events are comparatively rare. There is never a double-strand break and free ssDNA ends are held within the complex during the reaction, minimizing DNA damage responses.

Despite their potential, recombinases as a class are underdeveloped since most site-specific recombinases are not easily reprogrammable. The most highly active recombinases display a complex arrangement of DNA-binding and protein-protein contacts spread throughout

their structure (**Figure 3.2**),[17,18] and are consequently not amenable to simple truncation and translational fusion to more programmable DNA-binding domains.

Previous attempts at evolving recombinases have yielded enzymes with low activity,[19] or have required hundreds of rounds to achieve modest levels of retargeting.[20] To address this problem, we developed a genetic selection based on the Phage-Assisted Continuous Evolution

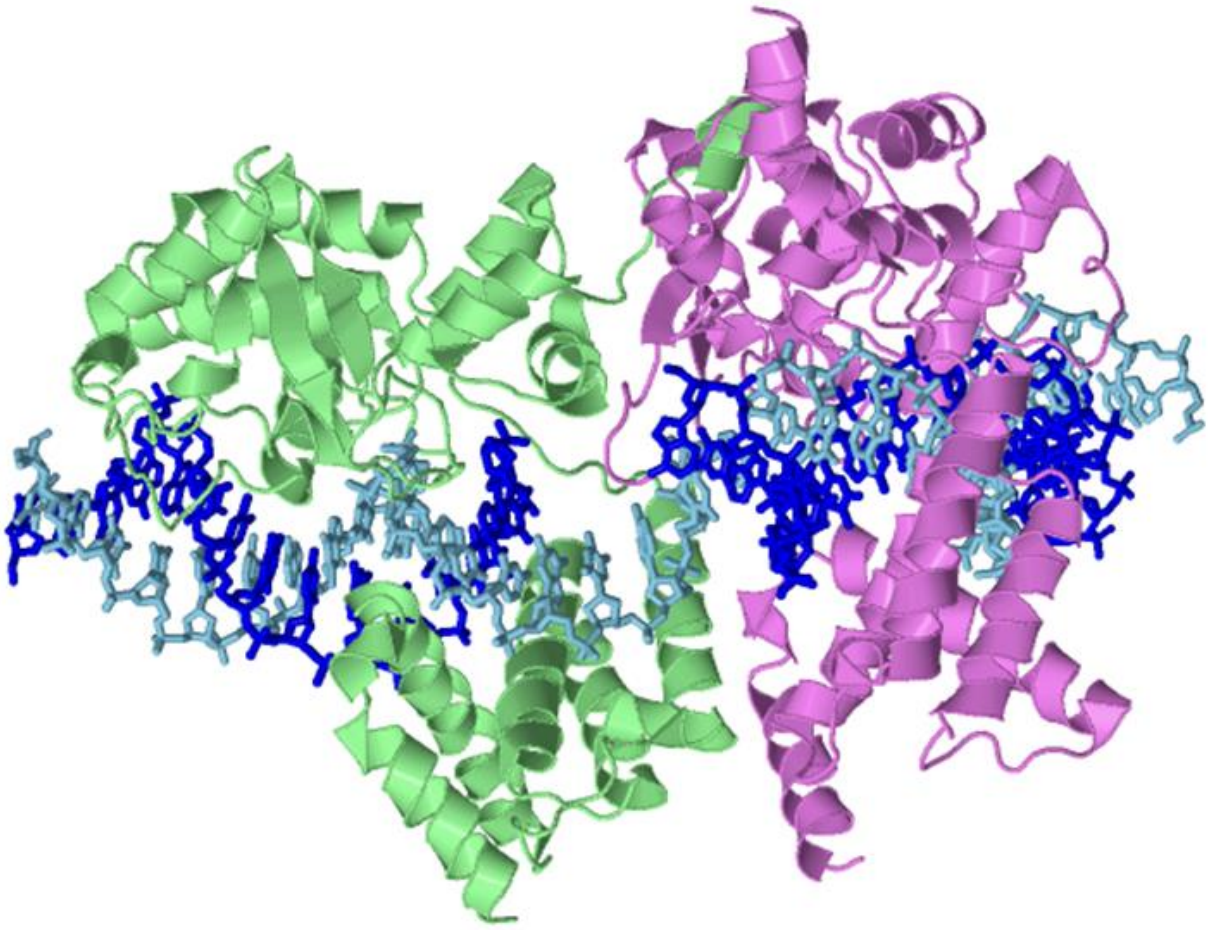

Figure 3.2 Recombinase complex assembly. Recombinases monomers (in this case, Cre) shown in green and purple bind to DNA over an extended surface area spread across both domains of the protein. Both N- and C-terminal lobes of the protein participate in dimerization interactions, as well (unpictured) tetramerization.

(PACE) system,[21] to enable the rapid evolution of recombinase proteins towards targets of interest. Using Cre recombinase as a model, the PACE system was optimized, validated, and used to evolve Cre variants with higher activity on their native loxP target site, as well as altered specificity towards a human genomic sequence within the hROSA26 locus.[22]

Results

PACE operates by linking the survival of desired functional genetic sequences to the filamentous phage life cycle. The system relies on the essential activity of the filamentous phage gene III which functions during both initial infection of host cells, and secretion of progeny phage from an infected host.[23] The entire selection takes place within a continuous flow apparatus, which we have currently adapted and simplified from previously published sytems (**Figure 3.3**).[21,24] A chemostat, receiving a constant influx of fresh media, sustains a constantly replicating population of uninfected host cells throughout the course of each PACE experiment. Cells from this chemostat are flowed into a separate vessel, termed the cellstat, where a population of phage particles reside, each carrying a copy of a 'selection phagemid' SP. Phage capable of infecting the host cells and activating expression of phage gene III from a so-called 'accessory plasmid' (AP) are able to produce infectious progeny that are secreted from the host cell and reinitiate the process on a fresh host cell. Phage that are incapable of eliciting gIII expression from the AP produce fewer or no progeny and will be diluted away under continuous flow. The flow rate in the cellstat is set such that there is no net-growth of host cells as their residence time is shorter than the average bacterial replication cycle. As the phage life cycle can be as short as 15 minutes, the only genetic material actively propagating within the system is that of the SP. This short cycle allows for dozens of rounds of selection, mutagenesis, and

amplification to happen in a single 24 hour period, potentially exploring more variants than traditional manual iterative directed evolution, and reaching more distant mutated states.

This cycling process takes place under mutagenesis, driven by a 'mutagenesis plasmid' (MP) that is induced upon entry into the cellstat to express mutagenic gene products. The MP carries three genes that augment mutagenic rate within the cells. A dominant negative form of the *E. coli* DNA polymerase proof-reading subunit dnaQ increases base misincorporation during phage genome replication.[25] Dam methylase is overexpressed to hypermethylate newly synthesized DNA and prevent the cell from discriminating parent strand from daughter strand during mismatch repair. Finally, the φ29 phage-derived protein p56 inhibits uracil DNA glycosylase,[26] resulting in an increase in C→T transitions.

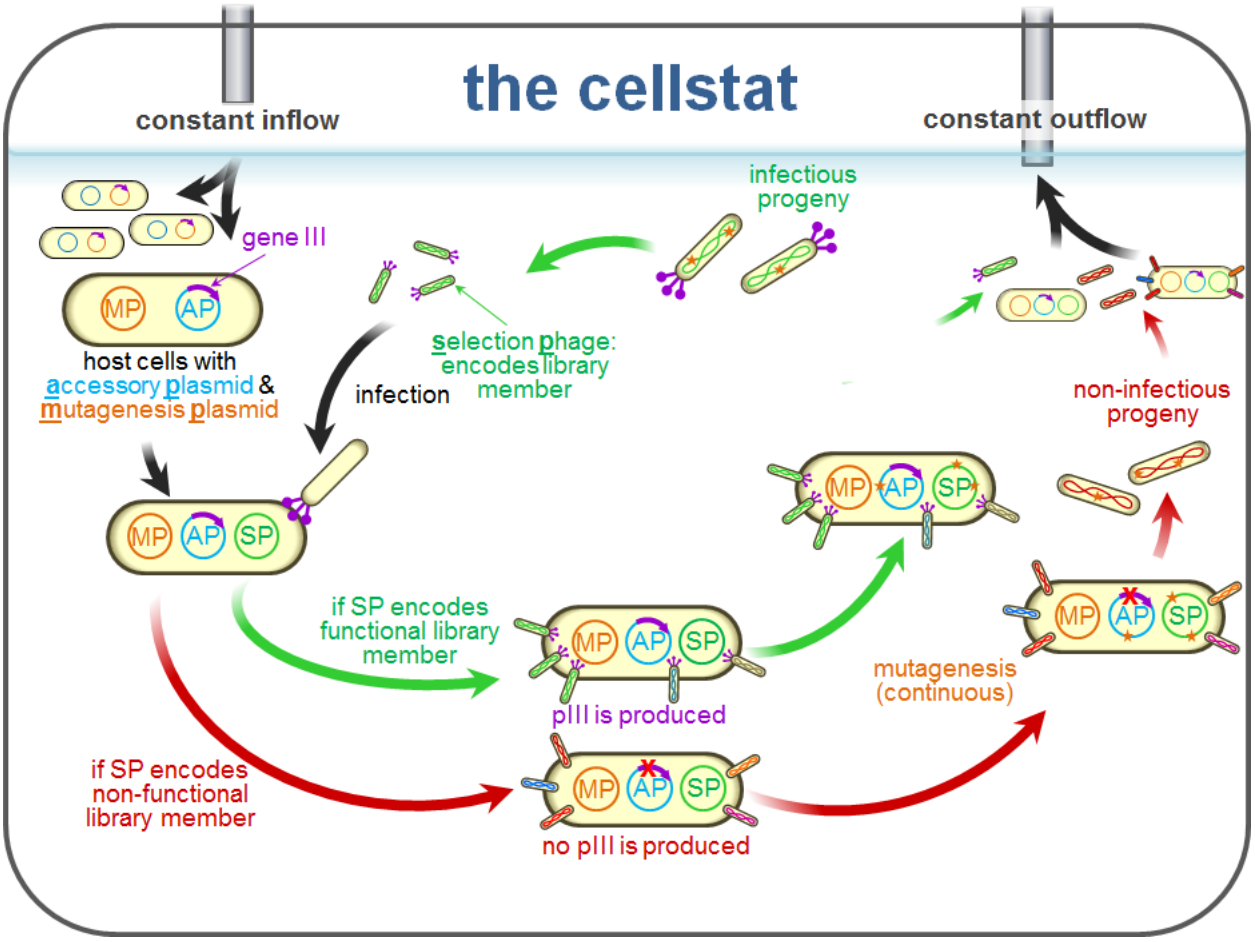

Figure 3.3 Phage-Assisted Continuous Evolution. The function of an evolving gene of interest is linked to the phage life cycle by having the gene encoded on the selection phagemid (SP) carryout an activity that elicits downstream gIII expression from an engineered accessory plasmid (AP). A mutagenesis plasmid (MP) provides constant mutagenesis. Active variants surive dilution relative to inactive variants.

To establish a recombinase-based genetic reporter system, we chose the tyrosine recombinase Cre as a model around which to develop the selection. Cre is an ideal candidate due to the wealth of biochemical and structural information available,[17,27] and it is among the most used, and most highly active recombianses in biological research. Cre binds to a specific 34-bp DNA sequence, called LoxP, and catalyzes recombination in a manner that is controlled by

religation of nicked strands post-exchange depending on specific hybridization of non-palindromic sticky ends within the recombination complex (**Figure 3.4**). This directional sequence, flanked by inverted repeats of monomeric Cre binding sites, is what determines the outcome of any given recombination reaction.[27,28]

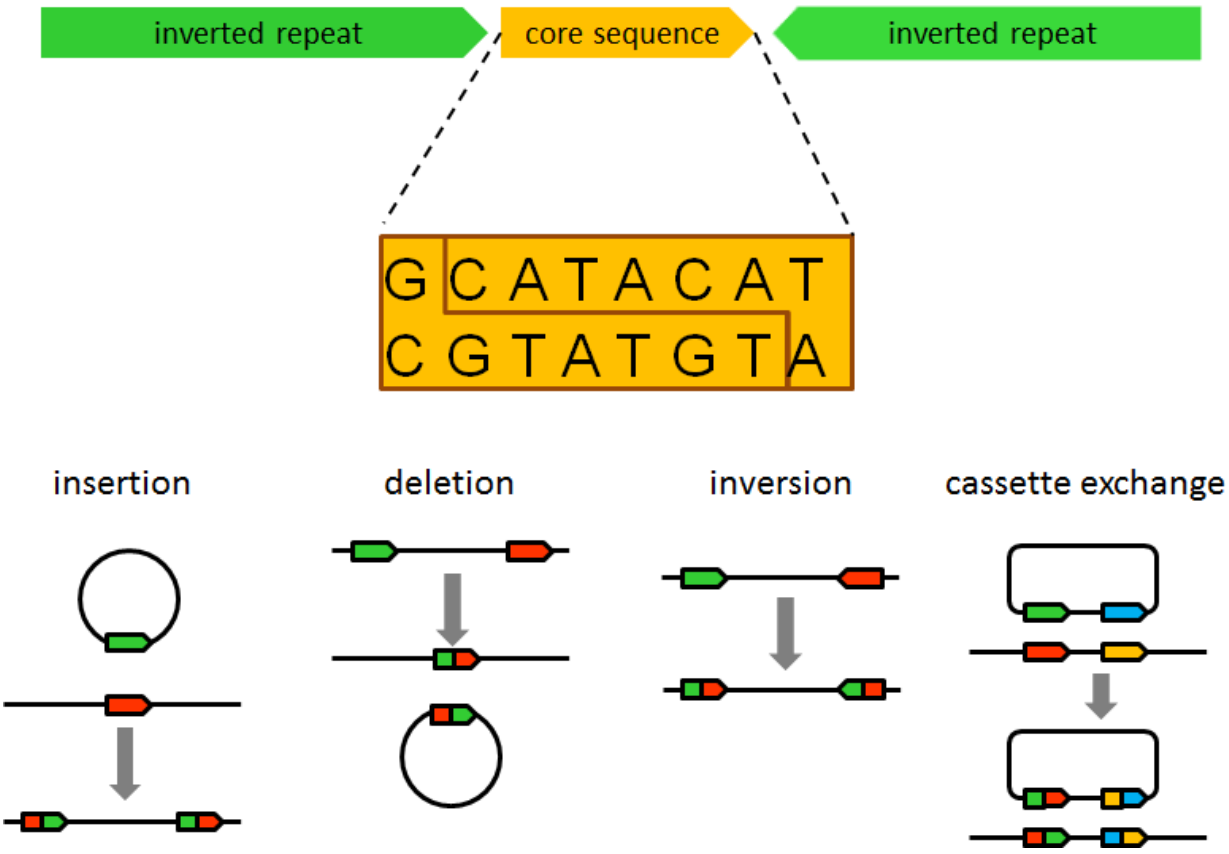

Figure 3.4 Directionality of Cre recombination is controlled by the hybridization and relegation of the core sequence. Cre nicks the core sequence on either side resulting in a 6 base 5' overhang that must anneal prior to ligation of exchanged strands. The orientation of this core sequences is what gives a LoxP binding site directionality, and this controls the progression and direction of recombination.

We assessed the effect of Cre-induced recombination on the expression of a luciferase reporter gene within either inversion or deletion based constructs. Inversion-based reporters

were constructed with a the luciferase coding sequence cloned in an inverted orientation relative to the upstream promoter, flanked by two inverted loxP sites. Deletion-based reporters were constructed by placing a strong transcriptional terminator flanked direct repeats of by loxP sites upstream of luciferase. In both constructs, expression through the recombination cassette was driven by an IPTG-regulated promoter, while Cre expression from a separate plasmid was driven by a tetracycline-regulated promoter.

We found that upon induction of expression through the recombination cassette, inversion-based reporters showed as much as 50% lower recombination rates over time (**Figure 3.5a**). This may be due to continued interaction between transcriptional machinery and recombination complexes, as the inversion reaction is completely reversible and equilibrates between inverted orientations. An alternative explanation involves expression of long antisense mRNA in the inverted orientation, which may interfere with expression o the luciferase reporter gene. Deletion based reporters showed a much lower effect of expression through the recombination cassette. This result was reflected in the level of luciferase signal produced by either deletion or inversion-based reporters (**Figure 3.5b**), with deletion reporters showing an increased signal concomitant with increased recombination over the same period of time (Figure M2.2A), while inversion reporters hit a ceiling of low activity after 3 hours, never reaching the level of activity seen with the deletion reporter. Given that rounds of PACE selection happen on the order of minutes, a reporter which manifests higher gene expression signals more quickly is the more desirable configuration. For this reason, we have used deletion-based reporters in all subsequent work.

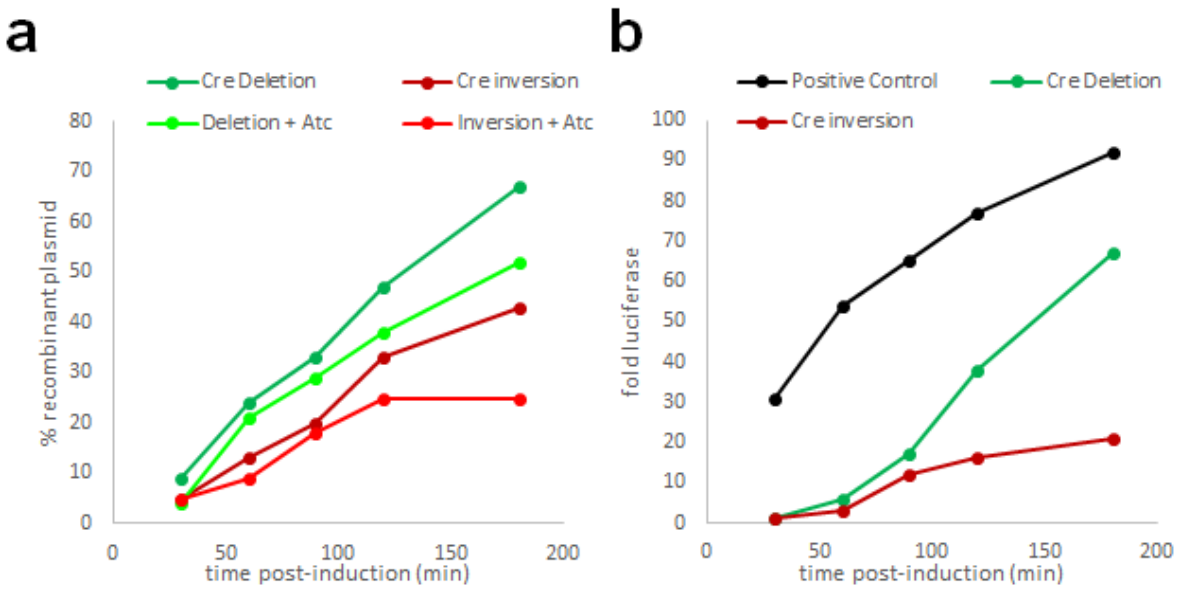

Figure 3.5 Performance of Cre inversion and deletion reporters. (**a**) interaction of gene expression an recombination on deletion- and inversion-based reporters, as measured by quantitation of a PCR-based gel electrophoresis assay of recombinant plasmids. (**b**) fold luminescence signal over uninduced controls for both deletion and inversion reporters. The positive control shown is the signal of a pre-deleted reporter plasmid derived from the deletion reporter.

Deletion AP constructs were prepared with the phage-shock promoter (PSP) upstream of a deletion cassette containing the rrnB transcriptional terminator flanked by loxP sites, with gene III and luciferase downstream (**Figure 3.6**). Infection with phage induces transcription from the PSP, which is terminated until recombinase-mediated deletion of rrnB enables expression of the downstream translationally-coupled gene III and luciferase reporters.

Cre-containing SP (Cre-SP) constructs were prepared and found to support activity-dependent phage production from recipient cells carrying the deletion AP in discrete overnight phage plaque assays. Based on this, we attempted continuous propagation across a range of dilution rates, and found that Cre-SP maintaining itself under continuous dilution rates as high as

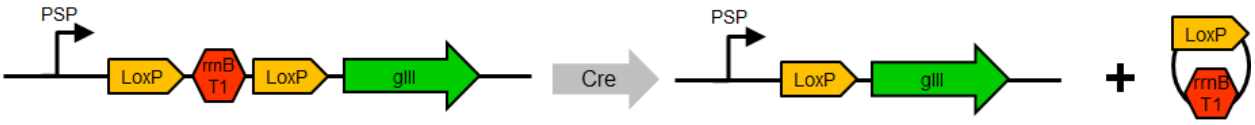

Figure 3.6 Layout of the Cre deletion-dependent AP construct. An rrnB terminator flanked by directed repeats of LoxP is positioned between the PSP and the gIII coding sequence. Successful deletion results in expression of gIII from the AP, leading to phage propagation.

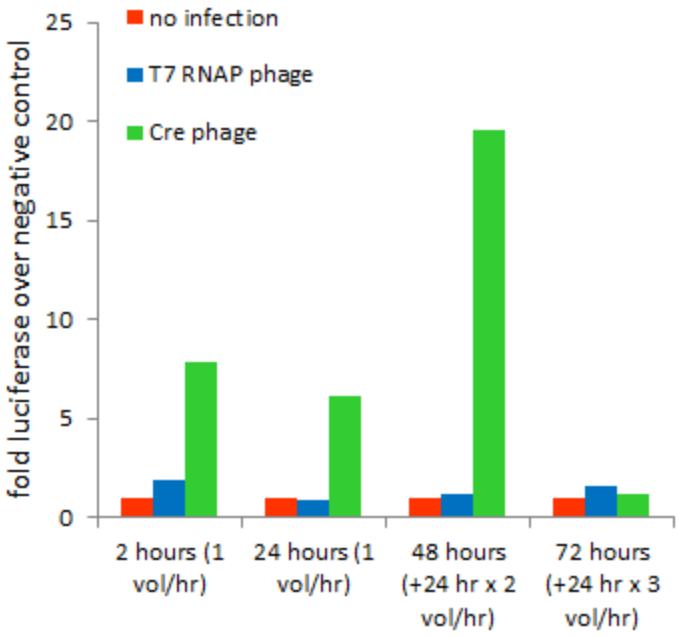

Figure 3.7 Continuous propagation of Cre-SP phage. Phage carrying Cre-SP, but not negative control T7 RNAP-SP, are capable of propagating and inducing coupled luciferase expression on deletion-dependent LoxP AP host cells at 2 vol/hr.

2 volumes per hour (**Figure 3.7**), comparable to our previously demonstrated T7 RNAP-bearing SPs which survive several days at 2 volumes per hour (vol/hr), but fail to propagate at 3 vol/hr, suggesting that Cre-SP may be similarly evolved within PACE.

Next, we performed a mock selection experiment by infecting a continuous culture of deletion AP-bearing recipient cells with a phage in a 1,000,000:1 ratio of negative control phage to Cre-SP phage. By 12 hours post-infection, Cre-SP had come to dominate the cellstat, with titers of ~10^8 pfu/mL, and propagating phage population converted completely to Cre-SP, with no detectable control phage present, as analyzed by a gel electrophoresis-based assay (**Figure 3.8**).

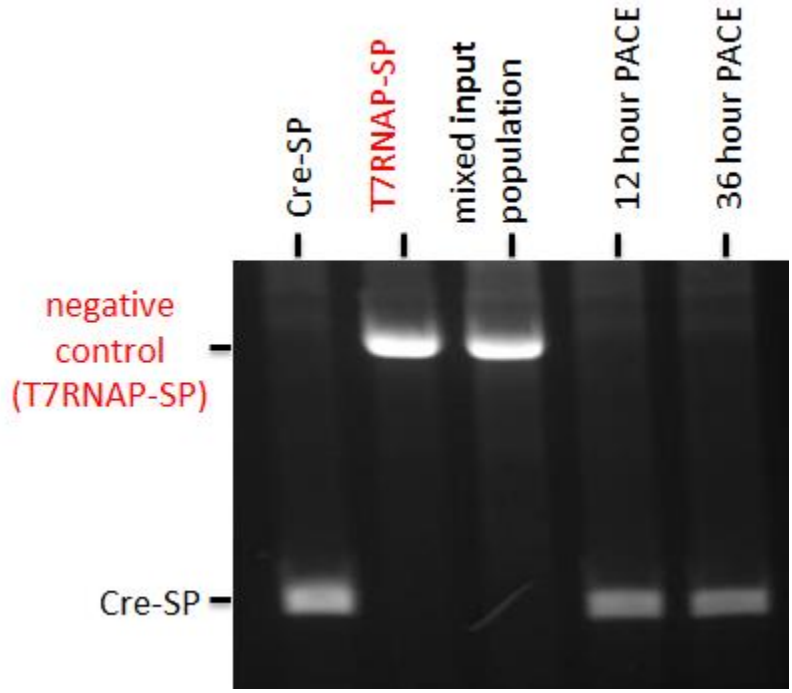

Figure 3.8 Mock enrichment of functional SP from an excess of non-functional SPs. PCR-based detection and gel electrophoresis of from the indicated phage populations and culture supernatants. PACE selection enriches Cre-SP from a 1:1x10^6 SP mixture in 12 hr.

We validated the ability of the system to arrive at working Cre variants from a starting catalytically inactive mutant clone (K201N).[29] Following 24 hours of propagation and mutagenesis with gIII freely provided on an expression plasmid (a process referred to as 'drift')[21,24] to allow phage survival from an initially inactive starting point, the phage population

was transferred to a new cellstat containing loxP.deletion AP host cells. After 24 hours, a population of functional Cre SPs had come to occupy the cellstat. Sequencing revealed that all clones had mutations at the inactivated position, most restoring the wild type residue, but others choosing an alternative (K201D). Luciferase activity assays showed that all surviving variants were active, whereas the starting variant was not (**Figure 3.9**).

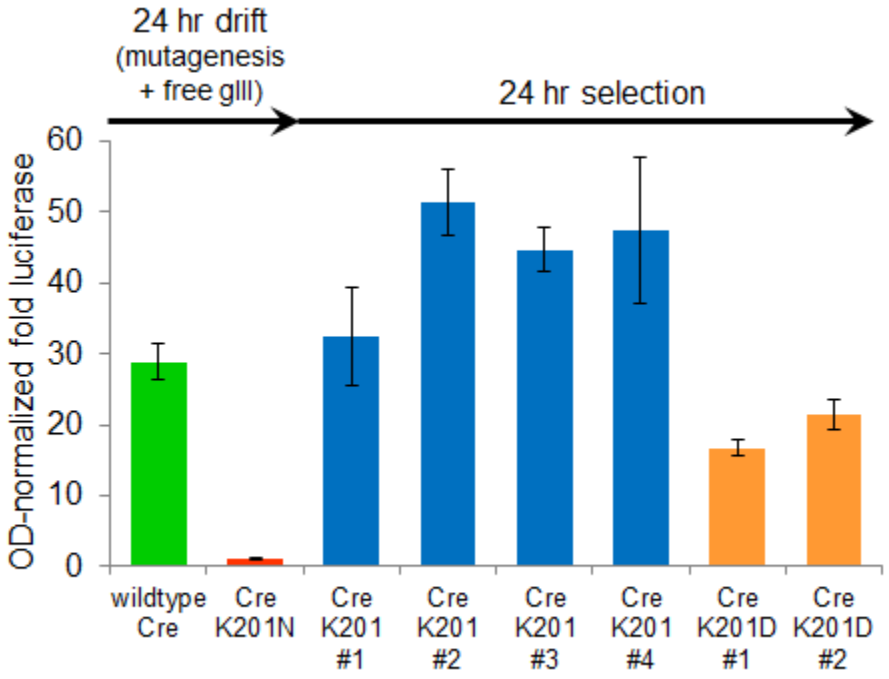

Figure 3.9 Mutagenesis and rescue of active variants from an inactive clonal starting point. 24 hours of drift followed by 24 hours of LoxP deletion selection resulted in the isolation of a mixed population of active variants.

As an initial target of Cre evolution, we identified several loxP-like sequences within the human ROSA26 locus[22] to evolve altered specificity towards (**Figure 3.10**). One of these sequences, designated ROSALoxP-7, was identified as having the fewest substitutions at

positions within the LoxP site known to be most important for Cre binding[27,30] and would thus be a potentially feasible target to reach.

wt	**AT**A**AC**TT**C**G**TATA** GCA**TA**CAT **TATA**C**G**AA**GT**T**AT**	0	(L:0 c:0 R:0)
1	CTAACTGGTAAGA ACAGAAAT GAGAGGTATTTAG	15	(L:6 c:3 R:6)
2	ATTCCTTCAAAAA GCTCAGAA TAGTTAAGGTAAT	15	(L:5 c:4 R:6)
3	TTATATACAAGGA ACATACTA AATACACAGGTAT	15	(L:8 c:3 R:4)
4	ATCCATTCTTAAA TGAAAAAA TATATAAATATTT	17	(L:5 c:5 R:7)
5	ATAAGTTACTTTA TTATACAC TAGAAACATACAC	15	(L:4 c:3 R:8)
6	AACATATAGTTTA GCATAAAC CATGAGATTTTAA	14	(L:6 c:2 R:6)
7	ATCTCATGGTTTA TGCTAAAC TATATGTTGACAT	15	(L:5 c:5 R:5)
8	GTTAAATCTTAAA CCCTACAG TACATAACTTTTT	16	(L:7 c:3 R:6)
9	ATAGGCCTGTATA CCAAATTT TAAACCATGTTGA	14	(L:5 c:4 R:5)

Figure 3.10 LoxP-like sequences within the human ROSA26 locus. Sequences found within the human ROSA26 locus with the fewest mutations from wild-type LoxP. Sequence 7 ('ROSALoxP-7') was chosen for a target due to the minimal number of base substitutions at sites of important protein:DNA contanct. In the top 'wt' line, the bold underlined positions correspond to the most important structural contacts. Red positions are substitutions away from the wild-type LoxP sequence.

As the ROSALoxP-7 sequence is asymmetric, we chose to split the evolution into two paths, one to evolved specificity of the left half-site and core sequences, and the other for the right half-site and core sequences (**Figure 3.11**). The resulting enzymes, active on their respective symmetric LoxP-like sites, would then be used as a heterodimer against the full asymmetric site. Given previous demonstrations of heterodimeric function of engineered Cre and other tyrosine recombinases, we expect our evolution strategy to be feasible.[20,31]

Figure 3.11 progression of evolutionary intermediate target sites for the left and right arcs towards ROSALoxP-7. Each evolutionary arc is split into four intermediates, substituting a single important structural contact (and additional less essential positions) at each step. Colors correspond to the stage at which the substitution is introduced.

Thus far we have carried forward the left arc of the ROSALoxP-7 evolution successfully through step 3 (7L3). At each step, the surviving population of SPs are passaged forward to initiate the evolution, which proceeds through a period of propagation on the current AP, a period of 1:1 mixing of two host cell AP populations in the cellstat, followed by a final propagation on the next AP alone. As of step 7L3, all clones have detectable activity on 7L3 APs, where wild-type Cre does not (**Figure 3.12**). Though the mixing transition strategy proved necessary for progression of the selection, one probably consequence of this is that promiscuous

variants dominate the population. This can be seen in the moderate but broad activity of the clones subjected to deletion-dependent luciferase assays in **Figure 3.12**. All clones retain activity on prior evolutionary intermediate target sequences.

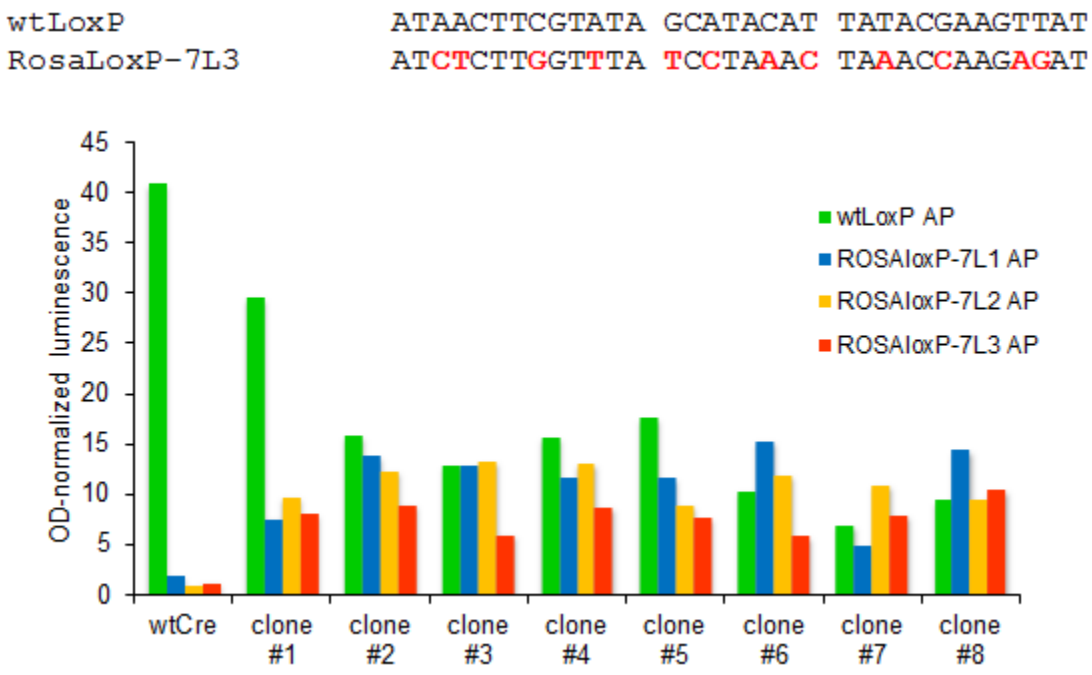

Figure 3.12 Activity of ROSAloxP-7L arc evolved population. All clones have gained appreciable activity on the 7L3 AP target, but retained promiscuity on prior intermediates.

Sequencing of phage clones has revealed the population converging on a number of consensus mutation positions. The 7L3-selected SP clones all contained 5 consensus mutations (M44V, A53E, A249V, R259C, and E262A), in addition to 1-3 non-shared coding mutations per clone. These 5 mutations are located within helices involved in protein:DNA contacts within the Cre recombinase crystal structure (**Figure 3.13**), and 3 of these mutations appear to be involved in direct contacts with the bases substituted in the mutated ROSALoxP intermediate DNA sequence (**Figure 3.14**).

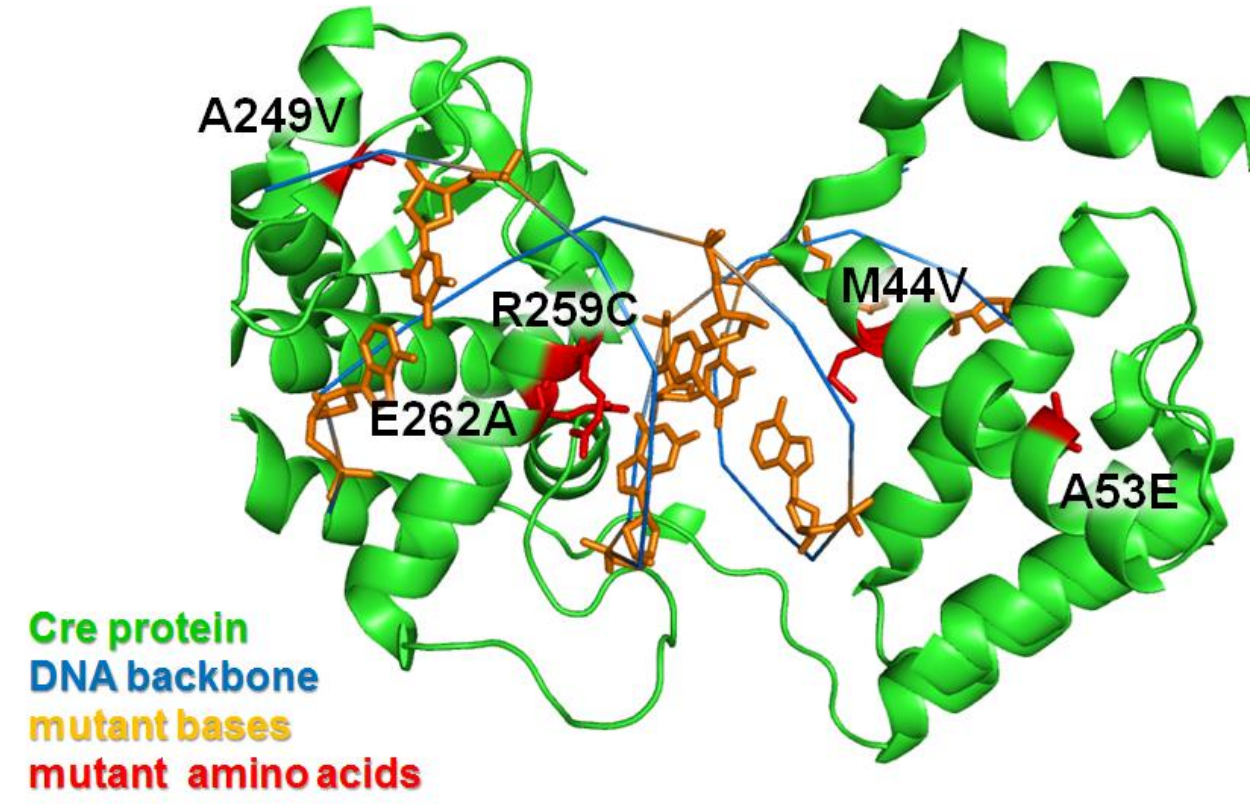

Figure 3.13 Converged mutations from the ROSAloxP-7L3 evolution map to helices and positions involved in DNA-binding and protein-protein interactions. Consensus positions in the 7L3 evolved population include M44V, A53E, A249V, E262A, R259C.

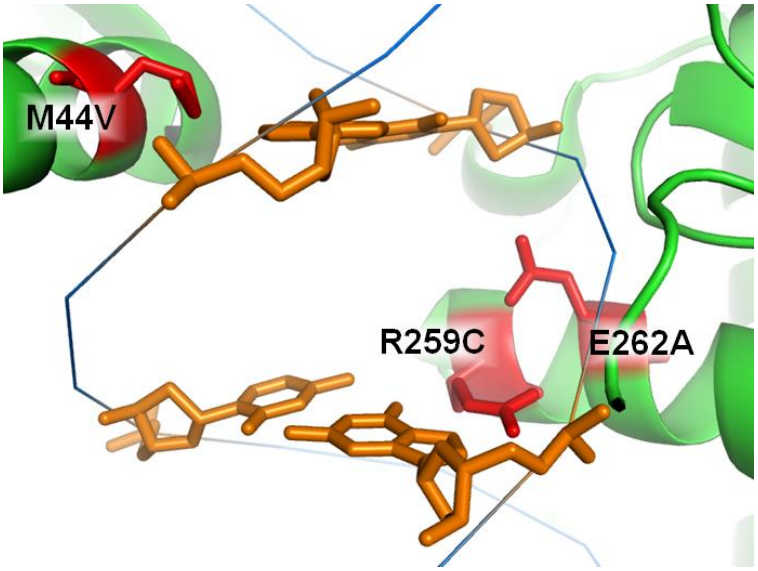

Figure 3.14 Three consensus positions make direct contact with substituted target base positions. Converged positions M44V, R259C, and E262A make direct contacts with substituted target bases.

We have also initiated evolution of Cre towards the right arc of the ROSALoxP-7 evolution, and have evolved a population in the same fashion as the left arc up through the intermediate target 7R2 (**Figure 3.11**). Once again, the population shows moderate levels of, broad deletion-dependent luciferase reporter activity on all intermediate sequences it has been exposed to, and have notably higher activity than the otherwise inactive wild-type Cre on these intermediate target sequences (**Figure 3.15**). Some clones show weak activity on 7L3 as well, suggesting that the next intermediate evolution has some viable starting activity. Sequencing of 7R2-evolved clones reveals a pattern of mutations similar to that of the 7L3 clones, with mutations present both at the site of protein-protein interactions and specific substituted base contacts (**Figure 3.16**). Consensus positions at E262G, I306V, I320M is common across almost all clones. Among the converged positions, only the E262 position can be seen as making a

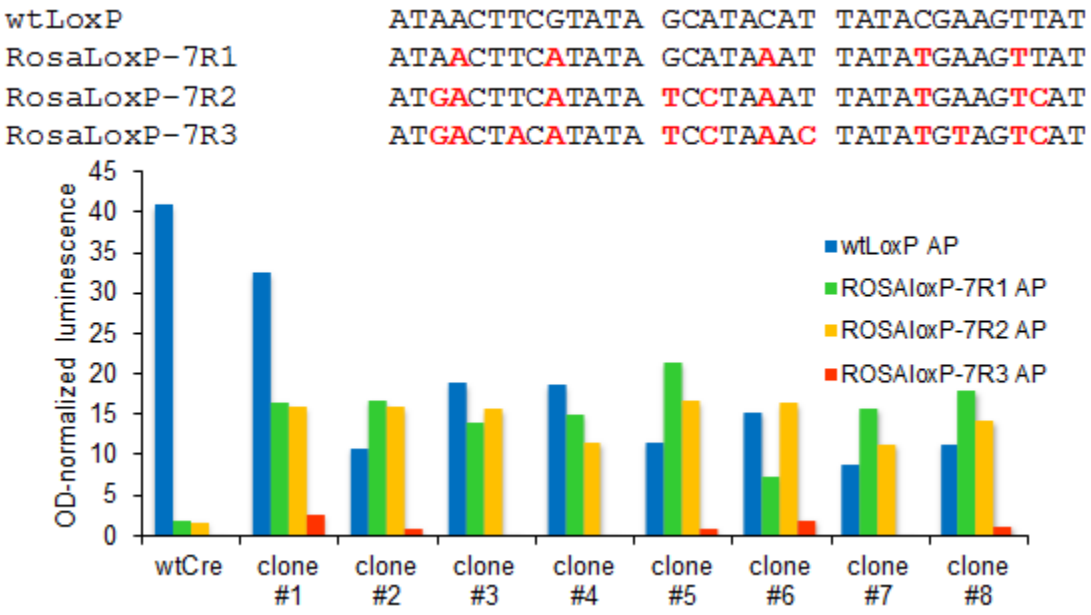

Figure 3.15 Activity of 7R2-evolved clones. All clones have gained appreciable activity on the 7R2 AP target, but retained promiscuity on prior intermediates.

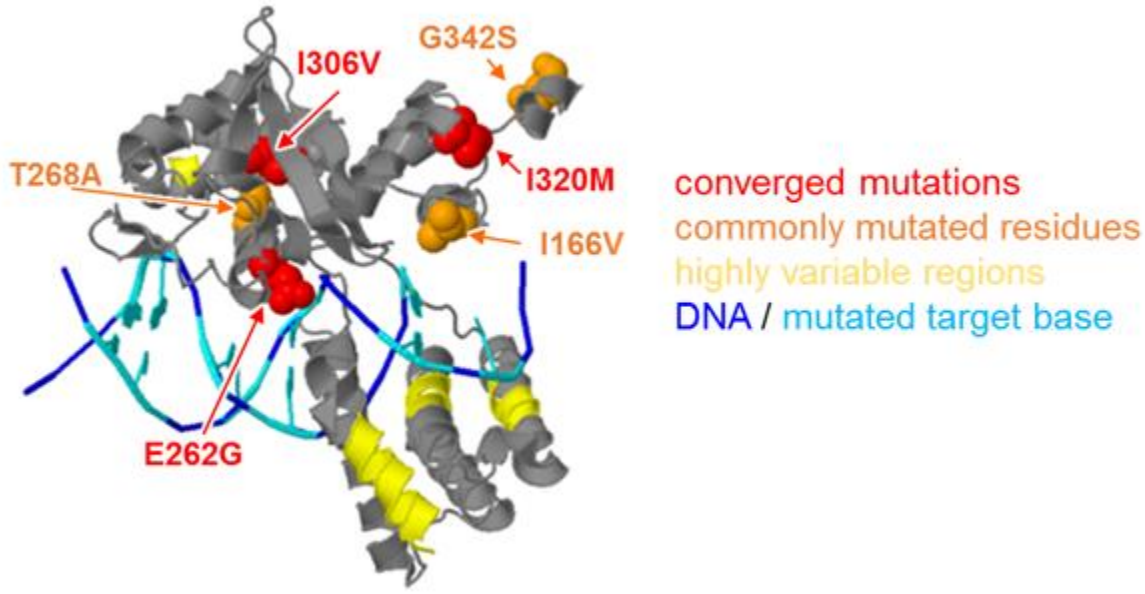

Figure 3.16 Consensus mutations in the ROSALoxP7-R arc as of 7R3. Converged positions at E262G, I306V, and I320M and additional repeating substitutions at I166V, T286A, and G342S

direct contact with one of the mutated target bases (**Figure 3.16**), and two distinct mutations (E262G/A GAA->GGA or ->GCA) appear to have arisen independently, suggesting a central role for this position in operating on the mutated R2 target sequence. Other converged positions are either internal to the protein (I306V), or located on a dimerization interface (I320M). Among the population containing the core triple-mutant set, there are some common and repeated, though not yet converged mutations, including I166V which can be seen to make contact with one of the target DNA bases. Finally, a large number of non-repeating mutations are located in the helices in the N-terminal half of the protein (the lower portion of the protein in **Figure 3.16**) that contact mutated bases in the target sequence. Together, this pattern of mutations suggests that our 7R arc evolution is selecting for Cre variants with functional changes at the DNA-binding interface, and not merely enhancing dimerization interactions or protein folding.

Finally, we have evolved Cre enzymes with apparent higher recombination rates in *E. coli* on the wild type loxP sequence. While Cre-based deletion is currently a powerful tool in basic research, synthetic biology, and circular DNA vector development, issues with the speed, efficiency, and occurrence of non-deletion side-reactions make the use of Cre less than ideal. For example, faster and more efficient Cre deletion variants may achieve higher knockout efficiency in 'floxed' transgenic mouse lines, resulting in more uniform conversion within the population of Cre-expressing cells. A faster Cre may improve dynamics in recombinase-dependent synthetic biological memory and counting applications[32]. Cre variants evolved to be more effective at deleting DNA between loxP sites maybe enable higher production yields of prokaryotic sequence-free minicircle DNA for use in DNA vaccines and episomal transgenesis.[33]

Towards these goals, we evolved more active Cre variants by propagating Cre recombinase on wild type loxP deletion APs and increasing the flow rate 0.5 vol/hr every 24 hours until reaching a rate of 4 vol/hr. Whereas wildtype Cre fails to propagate in PACE at dilution rates higher than 2.5 vol/hr (and shows a 100-fold drop in titer compared to 2 vol/hr), we successfully evolved a Cre population capable of maintaining itself at high titers (~10^8 pfu/mL) for over 24 hours at 4 vol/hr. Sequencing of surviving phage showed evidence of some sequence convergence, and the most common mutations (multiple unique substitutions of A285 to D, V or T, as well as a reoccurring triple mutant M31I/E151G/G281D) were located either internally, possibly affecting internal packing of the protein and dynamics during recombination, or on surface areas involved in protein-protein interactions between Cre monomers. Characterization of the activity of these enhanced Cre variants is the subject of future and ongoing research.

Discussion

Given the above progress on Cre evolution using the PACE recombinase system, we plan to continue evolution of towards the human ROSA26 integration target sequence. The promiscuity of the 7L3 and 7R2-evolved populations is of great concern if the evolved ROSALoxP-7 heterodimer pair is to be used effectively. To this end, we will be initiating negative selections against previous intermediate target sequences using the recently published negative selection system that makes use of a dominant-negative form of gIII to delivery a tunable fitness cost to variants displaying broad activity profiles.

Methods

PACE apparatus setup

The PACE apparatus used in this work is a derivation of the previously published chemostat-based (versus turbidostat) PACE system with custom modifications.[21,24] A 150 mL Erlenmeyer flask (VWR) is used for the chemostat rather than a 500 mL bioreactor vessel (Belco). The flask is stoppered with a 19/22 mm rubber septum. Media, vent, pressure, waste, and cellstat-out lines are serviced with disposable 22 ga and 16 ga 4 inch veterinary needles (AirTite). All connection tubing is 1/16 inch inner diameter clear PVC tubing. Cellstats are made with 50 mL Erlenmeyer flasks sealed with 14/22 mm rubber septa. Chemostat is maintained at 100 mL throughout the course of each experiment. All cellstats are run at 20 mL when propagating phage on single AP populations and 40 mL when propagating on mixed APs.

Mutagenesis and recombinant cheater control

Mutagenesis is induced from MPs using 5 mM final concentration arabinose drip from a syringe pump. The cellstats are maintained under 5 µg/mL kanamycin to combat the generation of recombinant cheater phage, which commonly recombine gIII into the SP using the upstream KanR gene for a 5' junction.

Accessory Plasmid cloning

Cloning of AP variants was performed using restriction digest of a PCR product of the p175dc wild-type LoxP deletion AP with XhoI and NheI. PCR of the terminator deletion cassette with compatible primers generating an insert containing one full LoxP variant site at the 3' end of the terminator cassette, and the vector containing the paired LoxP site on the 5' end of the deletion cassette insertion site. All PCRs are digested for 1 hour, cleaned using Qiagen PCR cleanup kit, ligated for 10 minutes using the Quick Ligation kit (NEB), and transformed into Mach1 chemically competent cells (Life Technologies).

References

1. Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat. Protoc.* **4,** 1471–1501 (2009).

2. Reyon, D. *et al.* Engineering customized TALE nucleases (TALENs) and TALE transcription factors by fast ligation-based automatable solid-phase high-throughput (FLASH) assembly. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **Chapter 12,** Unit 12.16 (2013).

3. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).

4. Tebas, P. *et al.* Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. *N. Engl. J. Med.* **370,** 901–910 (2014).

5. Hartlerode, A. J. & Scully, R. Mechanisms of double-strand break repair in somatic mammalian cells. *Biochem. J.* **423,** 157–168 (2009).

6. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle Georget. Tex* **7,** 2902–2906 (2008).

7. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–84 (2014).

8. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16,** 2164–2173 (1996).

9. Ramirez, C. L. *et al.* Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. *Nucleic Acids Res.* **40,** 5560–5568 (2012).

10. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154,** 1380–1389 (2013).

11. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–651 (2005).

12. Maeder, M. L. *et al.* Rapid 'Open-Source' Engineering of Customized Zinc-Finger Nucleases for Highly Efficient Gene Modification. *Mol. Cell* **31,** 294–301 (2008).

13. Orii, K. E., Lee, Y., Kondo, N. & McKinnon, P. J. Selective utilization of nonhomologous end-joining and homologous recombination DNA repair pathways during nervous system development. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 10017–10022 (2006).

14. Saintigny, Y. *et al.* Characterization of homologous recombination induced by replication inhibition in mammalian cells. *EMBO J.* **20,** 3861–3870 (2001).

15. Grindley, N. D. F., Whiteson, K. L. & Rice, P. A. Mechanisms of Site-Specific Recombination*. *Annu. Rev. Biochem.* **75,** 567–605 (2006).

16. Janbandhu, V. C., Moik, D. & Fässler, R. Cre recombinase induces DNA damage and tetraploidy in the absence of LoxP sites. *Cell Cycle Georget. Tex* **13,** 462–470 (2014).

17. Guo, F., Gopaul, D. N. & van Duyne, G. D. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **389,** 40–46 (1997).

18. Guo, F., Gopaul, D. N. & Van Duyne, G. D. Asymmetric DNA bending in the Cre-loxP site-specific recombination synapse. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 7143–7148 (1999).

19. Gaj, T., Mercer, A. C., Sirk, S. J., Smith, H. L. & Barbas, C. F., 3rd. A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res.* **41,** 3937–3946 (2013).

20. Sarkar, I., Hauber, I., Hauber, J. & Buchholz, F. HIV-1 proviral DNA excision using an evolved recombinase. *Science* **316,** 1912–1915 (2007).

21. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472,** 499–503 (2011).

22. Irion, S. *et al.* Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat. Biotechnol.* **25,** 1477–1482 (2007).

23. Rakonjac, J. in *eLS* (John Wiley & Sons, Ltd, 2001). at <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0000777/abstract>

24. Carlson, J. C., Badran, A. H., Guggiana-Nilo, D. A. & Liu, D. R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat. Chem. Biol.* **10,** 216–222 (2014).

25. Foster, P. L., Gudmundsson, G., Trimarchi, J. M., Cai, H. & Goodman, M. F. Proofreading-defective DNA polymerase II increases adaptive mutation in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 7951–7955 (1995).

26. Asensio, J. L. *et al.* Novel dimeric structure of phage ?29-encoded protein p56: insights into uracil-DNA glycosylase inhibition. *Nucleic Acids Res.* **39,** 9779–9788 (2011).

27. Ghosh, K. & Van Duyne, G. D. Cre-loxP biochemistry. *Methods San Diego Calif* **28,** 374–383 (2002).

28. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450,** 56–62 (2007).

29. Gibb, B. *et al.* Requirements for catalysis in the Cre recombinase active site. *Nucleic Acids Res.* **38,** 5817–5832 (2010).

30. Thyagarajan, B., Guimarães, M. J., Groth, A. C. & Calos, M. P. Mammalian genomes contain active recombinase recognition sites. *Gene* **244,** 47–54 (2000).

31. Shaikh, A. C. & Sadowski, P. D. Chimeras of the Flp and Cre recombinases: tests of the mode of cleavage by Flp and Cre. *J. Mol. Biol.* **302,** 27–48 (2000).

32. Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol.* **31,** 448–452 (2013).

33. Chen, Z.-Y., He, C.-Y., Ehrhardt, A. & Kay, M. A. Minicircle DNA Vectors Devoid of Bacterial DNA Result in Persistent and High-Level Transgene Expression in Vivo. *Mol. Ther.* **8,** 495–500 (2003).

Chapter 4:

Improvement of Genome modification Specificity by Fusion of Inactivated Cas9 to *Fok*I Nuclease

Abstract

Programmable site-specific endonucleases have proven to be useful tools for genome editing and may lead to novel therapeutics to treat genetic diseases. Cas9 cleaves double-stranded DNA in a variety of organisms at a sequence programmed by a short single-guide RNA (sgRNA). The specificity of Cas9-mediated DNA cleavage is imperfect, and off-target cleavage both in the test tube and in cells has been observed.[1–5] Recently engineered variants of Cas9 that cleave only one DNA strand ("nickases") enable double-stranded breaks to be specified by two distinct sgRNA sequences,[5–7] but still suffer from off-target cleavage activity[7,8] arising from the ability of each monomeric nickase to remain active when individually bound to DNA.[9–11] Here we describe the development of a *Fok*I nuclease fusion to a catalytically dead Cas9 that requires simultaneous DNA binding and association of two *Fok*I-dCas9 monomers to cleave DNA. Off-target DNA cleavage of the engineered *Fok*I-dCas9 (fCas9) is further reduced by the requirement that only sites flanked by two sgRNAs ~15 or 25 base pairs apart are cleaved, a much more stringent spacing requirement than nickases. In human cells, fCas9 modified target DNA sites with efficiency comparable to that of nickases and with > 140-fold higher specificity than wild-type Cas9. At loci with highly similar off-target sites, fCas9 modified genomic DNA in human cells with greater specificity than Cas9 nickases. Target sites that conform to the substrate requirements of fCas9 are abundant in the human genome, occurring on average once every 34 bp.

Introduction

The recent development of robust, predictable, and user-friendly methods for the generation of sequence-specific DNA-binding proteins has led to a rapid expansion of the field

of genome editing. Today, user-defined site-specific genome modification has become a powerful tool in biological research[12], and holds significant potential to serve as the basis of a new generation of human therapeutics[13] (NCT00842634, NCT01044654, NCT01252641).

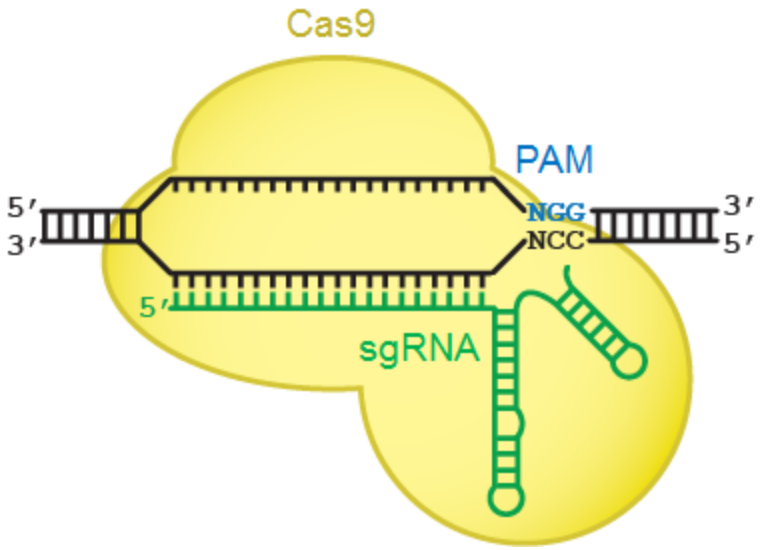

Figure 4.1 Architecture of Cas9. (a) Cas9 protein (yellow) binds to target DNA in complex with a guide RNA (sgRNA, green). The *S. pyogenes* Cas9 protein recognizes the PAM sequence NGG (blue), initiating unwinding of dsDNA and sgRNA:DNA base pairing.

One such programmable endonuclease system uses the CRISPR-derived Cas9 nuclease complexed with an engineered single-guide RNA (sgRNA) to target dsDNA sequences for cleavage (**Figure 4.1**).[10] The 3' end of the sgRNA forms a scaffold that binds Cas9 protein,[14] while the 5' most ~17 to 20 bases[8] of the sgRNA pair with the target DNA to determine DNA cleavage specificity. Provided that the target sequence is adjacent to a short 3' motif, the protospacer adjacent motif (PAM) required for initial binding and Cas9 activation[10] (**Figure 4.1**), any DNA locus can in principle be targeted. In cells, Cas9:sgRNA-induced double strand breaks can result in functional gene knockout through non-homologous end joining (NHEJ) or alteration of a target locus to virtually any sequence through homology-directed repair (HDR) with an

exogenous DNA template.[9,14,15] Cas9 is an especially convenient genome editing platform,[16] as a genome editing agent for each new target site of interest can be accessed by simply generating the corresponding sgRNA. This approach has been widely used to create targeted knockouts and gene insertions in cells and model organisms, and has also been recognized for its potential therapeutic relevance.

While Cas9:sgRNA systems provide a high level of programmability and ease of use, our group[1] and others[2–5] have reported the ability of Cas9 to cleave off-target genomic sites, resulting in modification of unintended loci that can limit the usefulness of Cas9 as a research tool and as a potential therapeutic. We hypothesized that engineering Cas9 variants to cleave DNA only when two simultaneous, adjacent Cas9:DNA binding events take place could substantially improve genome editing specificity since the likelihood of two adjacent off-target binding events is much smaller than the likelihood of a single off-target binding event (approximately $1/n^2$ vs. $1/n$). Such an approach is distinct from the recent development of mutant Cas9 proteins that cleave only a single strand of dsDNA ("nickases"). Nickases can be used to nick opposite strands of two nearby target sites, generating what is effectively a double strand break, and paired Cas9 nickases can effect substantial on-target DNA modification with reduced off-target modification.[5,7,8] Because each of the component Cas9 nickases remains catalytically active[9–11] and single-stranded DNA cleavage events are weakly mutagenic,[17,18] nickases can induce genomic modification even when acting as monomers.[5,6,14] Indeed, Cas9 nickases have been previously reported to induce off-target modifications in cells.[7,8] Moreover, since paired Cas9 nickases can efficiently induce dsDNA cleavage-derived modification events when bound up to ~100 bp apart,[6,7] the statistical number of potential off-target sites for paired nickases is larger than that of a more spatially constrained dimeric Cas9 cleavage system.

To further improve the specificity of the Cas9:sgRNA system, we sought to engineer an obligate dimeric Cas9 system analogous to previously developed dimeric zinc-finger nucleases (ZFNs) and TALENs. These nucleases have been widely used as research tools in cell culture and *in vivo*[19], and ZFNs are currently in clinical trials as potential human therapeutics.[13] Based on ZFN and TALEN examples, we speculated that fusing the *Fok*I restriction endonuclease cleavage domain to a catalytically dead Cas9 (dCas9) could create an obligate dimeric Cas9 that would cleave DNA only when two distinct *Fok*I-dCas9:sgRNA complexes bind to adjacent sites ("half-sites") with particular spacing constraints (**Figure 4.2**).

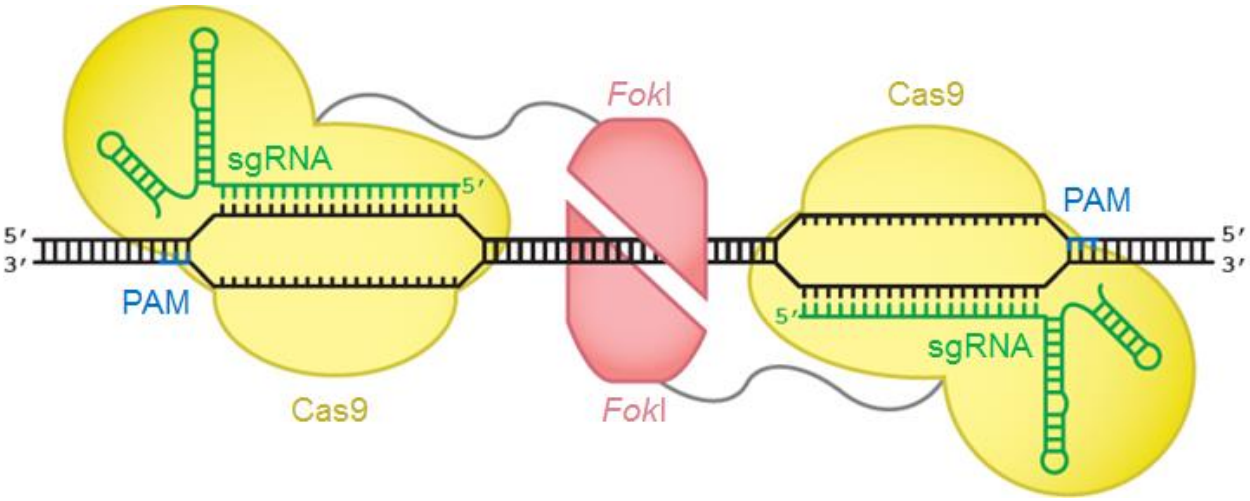

Figure 4.2 Architecture *Fok*I-dCas9 fusion variants. Monomers of *Fok*I nuclease (red) fused to dCas9 bind to separate sites within the target locus. Only adjacently bound *Fok*I-dCas9 monomers can assemble a catalytically active *Fok*I nuclease dimer, triggering dsDNA cleavage.

In contrast with Cas9 nickases, in which single-stranded DNA cleavage by monomers takes place independently, the DNA cleavage of *Fok*I-dCas9 requires simultaneous binding of two distinct *Fok*I-dCas9 monomers because monomeric *Fok*I nuclease domains are not catalytically competent.[20] In principle this approach should increase the specificity of DNA cleavage relative to wild-type Cas9 by doubling the number of specified target bases contributed

by both monomers of the *Fok*I-dCas9 dimer, and should also offer improved specificity compared to nickases due to inactivity of monomeric *Fok*I-dCas9:sgRNA complexes, and due to the more stringent spatial requirements for assembly of a *Fok*I-dCas9 dimer.

Results

While fusions of Cas9 to short functional peptide tags have been described to enable sgRNA-programmed transcriptional regulation,[21] to our knowledge no fusions of Cas9 with active enzyme domains have been previously reported. Therefore we began by constructing and characterizing a wide variety of *Fok*I-dCas9 fusion proteins with distinct configurations of a *Fok*I nuclease domain, dCas9 containing inactivating mutations D10A and H840A, and a nuclear localization sequence (NLS). We fused wild-type, homodimeric, *Fok*I to either the N- or C-terminus of dCas9, and varied the location of the NLS to be at either terminus or between the two domains (**Figure 4.3**). We further varied the length of the linker sequence as either one or three repeats of Gly-Gly-Ser (GGS) between the *Fok*I and dCas9 domains.

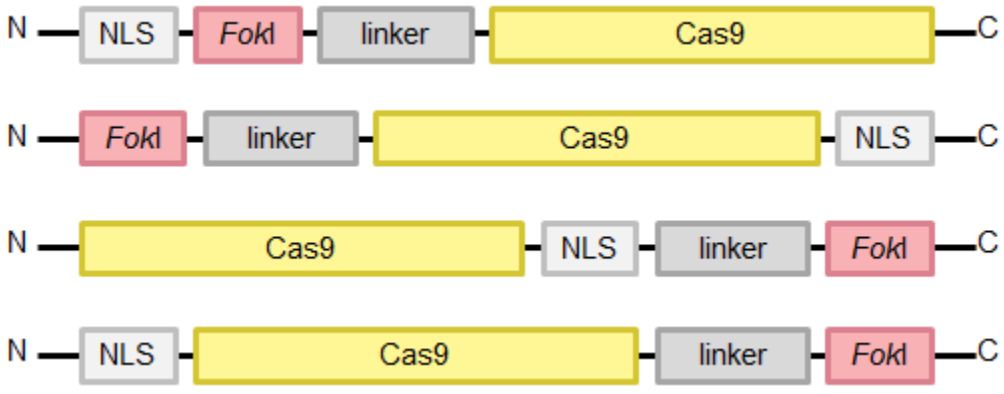

Figure 4.3 *Fok*I-dCas9 fusion architectures tested. Four distinct configurations of NLS, *Fok*I nuclease, and dCas9 were assembled, with the *Fok*I-dCas9 linker varied as either one or three repeats of (Gly-Gly-Ser) per configuration.

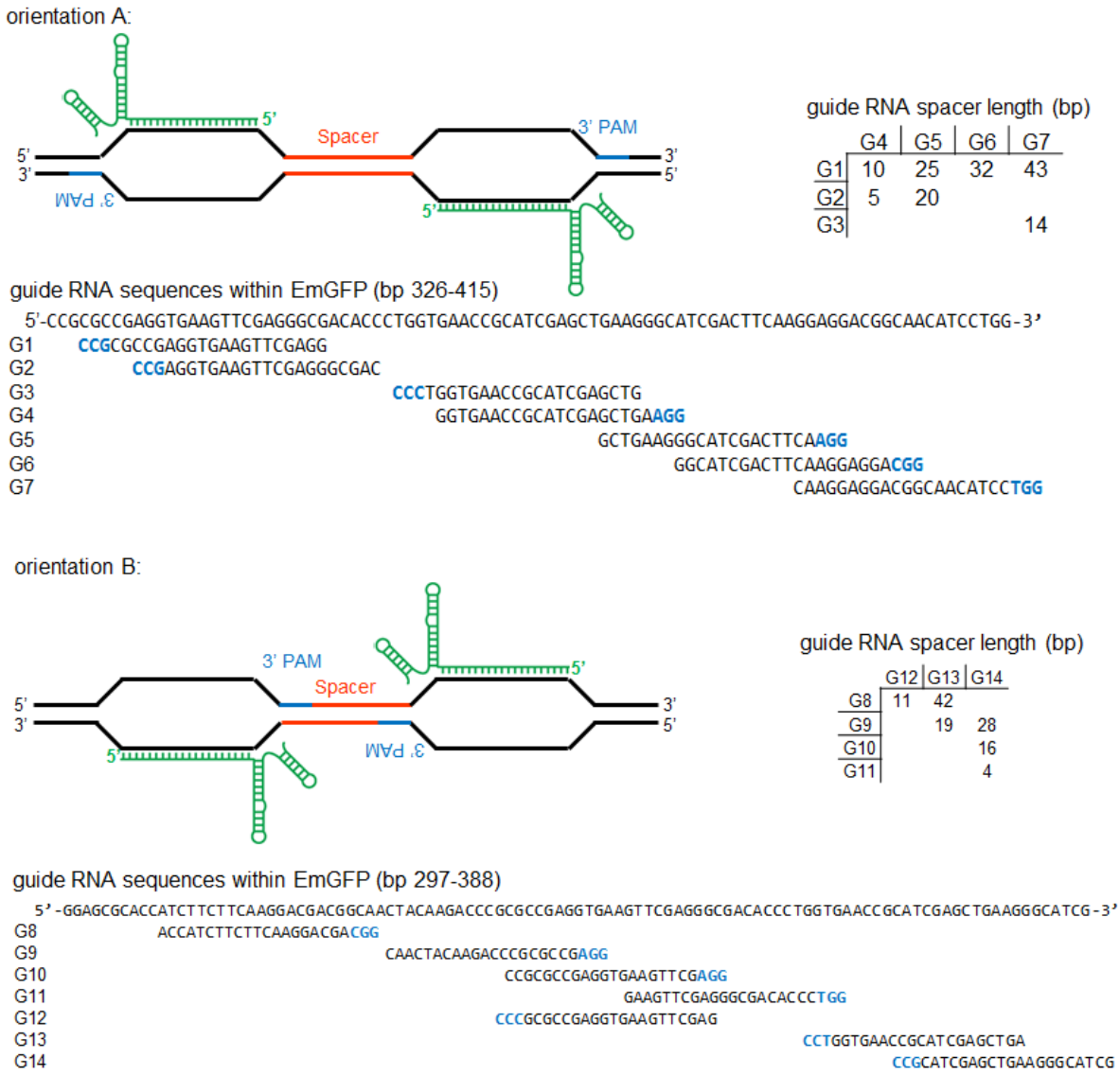

Figure 4.4 sgRNA target sites tested within EmGFP. Target sites were chosen to test *Fok*I-dCas9 activity either an orientation in which each PAM is distal from the cleaved spacer sequence (orientation A), or proximal to the cleaved spacer (orientation B). Together, these seven sgRNAs enabled testing of *Fok*I-dCas9 fusion variants across seven spacer lengths ranging from 5 to 43 bp in orientation A and six spacer lengths ranging from 4 to 42 in orientation B.

Since previously developed dimeric nuclease systems are sensitive to the length of the spacer sequence between half-sites,[22,23] we also tested a wide range of spacer sequence lengths between two sgRNA binding sites within a test target gene, Emerald GFP (Life Technologies) (referred to hereafter as GFP) (**Figure 4.4**). Two sets of sgRNA binding-site pairs with different

orientations were chosen within GFP. One set placed the pair of NGG PAM sequences distal from the spacer sequence, with the 5' end of the sgRNA adjacent to the spacer (orientation A),, while the other placed the PAM sequences immediately adjacent to the spacer (orientation B). In total, seven pairs of sgRNAs were suitable for orientation A, and six were suitable for orientation B. By pairwise combination of the sgRNA targets, we tested spacer lengths in both dimer orientations, ranging from 5 to 43 bp in orientation A, and 4 to 42 bp in orientation B. In total, 216 pairs of *Fok*I-dCas9:sgRNA complexes were generated and tested, exploring four fusion architectures, 17 protein linker variants (described below), both sgRNA orientations, and 13 spacer lengths between half-sites.

To assay initially the activities of these candidate *Fok*I-dCas9:sgRNA pairs, we used a previously described flow cytometry-based fluorescence assay[2,8] in which DNA cleavage and NHEJ of a stably integrated constitutively expressed GFP gene in HEK293 cells leads to loss of cellular fluorescence (**Figure 4.5**). For comparison, we assayed the initial set of *Fok*I-dCas9 variants side-by-side with the corresponding Cas9 nickases and wild-type Cas9 in the same expression plasmid across both sgRNA spacer orientation sets A and B. Cas9 protein variants and sgRNA were generated in cells by transient co-transfection of the corresponding Cas9 protein expression plasmids together with the appropriate pair of sgRNA expression plasmids. The *Fok*I-dCas9 variants, nickases, and wild-type Cas9 all targeted identical DNA sites using identical sgRNAs. While this assay showed a consistent ~5% background signal in the absence of sgRNA, it enabled the rapid assessment of many fusion constructs, sgRNA orientations, and DNA spacer lengths to identify active constructs for further optimization.

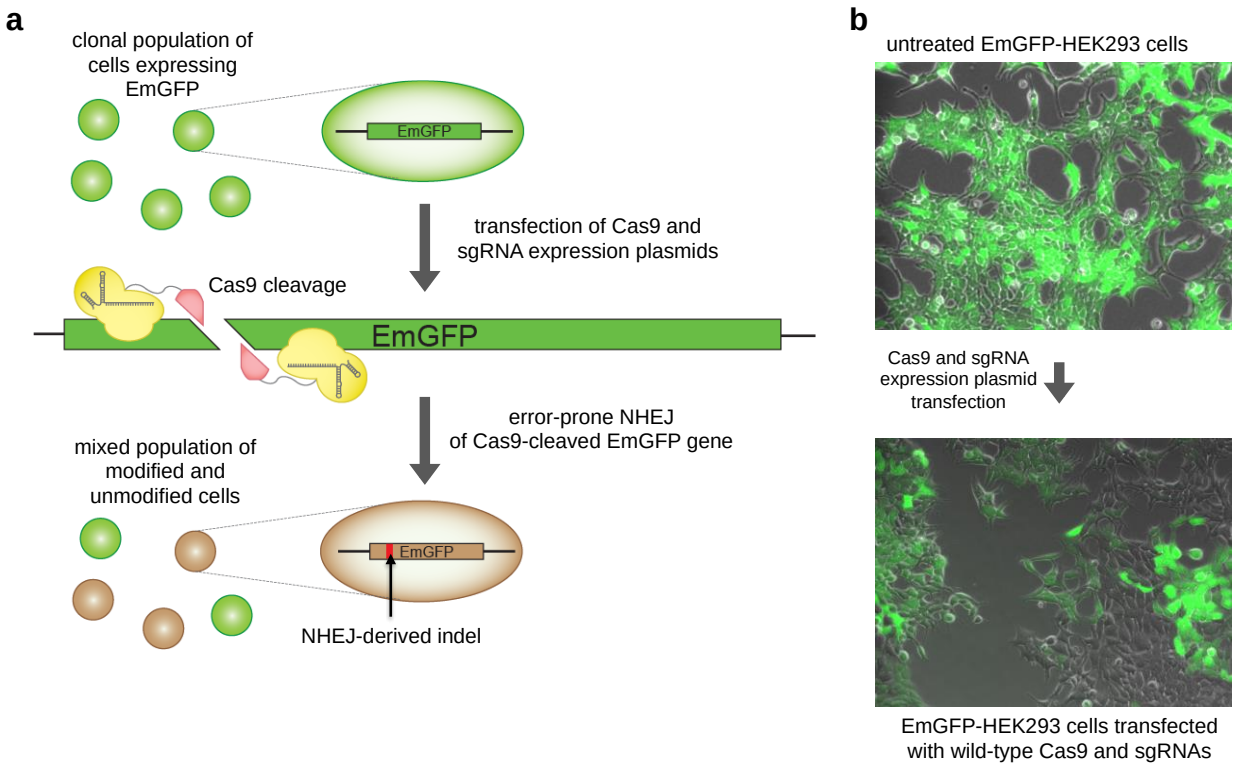

Figure 4.5. GFP disruption assay for measuring genomic DNA-modification activity. (**a**) A HEK293-derived cell line constitutively expressing a genomically integrated EmGFP gene was used to test the activity of candidate *Fok*I-dCas9 fusion constructs. Co-transfection of these cells with appropriate nuclease and sgRNA expression plasmids leads to dsDNA cleavage within the EmGFP coding sequence, stimulating error-prone NHEJ and generating indels that can disrupt the expression of GFP, leading to loss of cellular fluorescence. The fraction of cells displaying a loss of GFP fluorescence is then quantitated by flow cytometry. (**b**) Typical epifluorescence microscopy images at 200x magnification of EmGFP-HEK293 cells before and after co-transfection with wild-type Cas9 and sgRNA expression plasmids.

Most of the initial *Fok*I-dCas9 fusion variants were inactive or very weakly active (**Figure 4.6**). The NLS-*Fok*I-dCas9 architecture (listed from N to C terminus), however, resulted in a 10% higher frequency of GFP-negative cells above that of the corresponding no-sgRNA control when used in orientation A, with PAMs distal from the spacer. (**Figure 4.6a**). In contrast, NLS-*Fok*I-dCas9 activity above background was not detected when used with sgRNA pairs in orientation B, with PAMs adjacent to the spacer (**Figure 4.6b**). Examination of the recently reported Cas9 structures[24,25] reveals that the Cas9 N-terminus protrudes from the RuvC

domain, which contacts the 5' end of the sgRNA:DNA duplex. We speculate that this arrangement places an N-terminally fused *Fok*I distal from the PAM, resulting in a preference for sgRNA pairs with PAMs distal from the cleaved spacer (orientation A, **Figure 4.4**). While other *Fok*I-dCas9 fusion pairings and the other sgRNA orientation in some cases showed modest activity, we chose NLS-*Fok*I-dCas9 with sgRNAs in orientation A for further development.

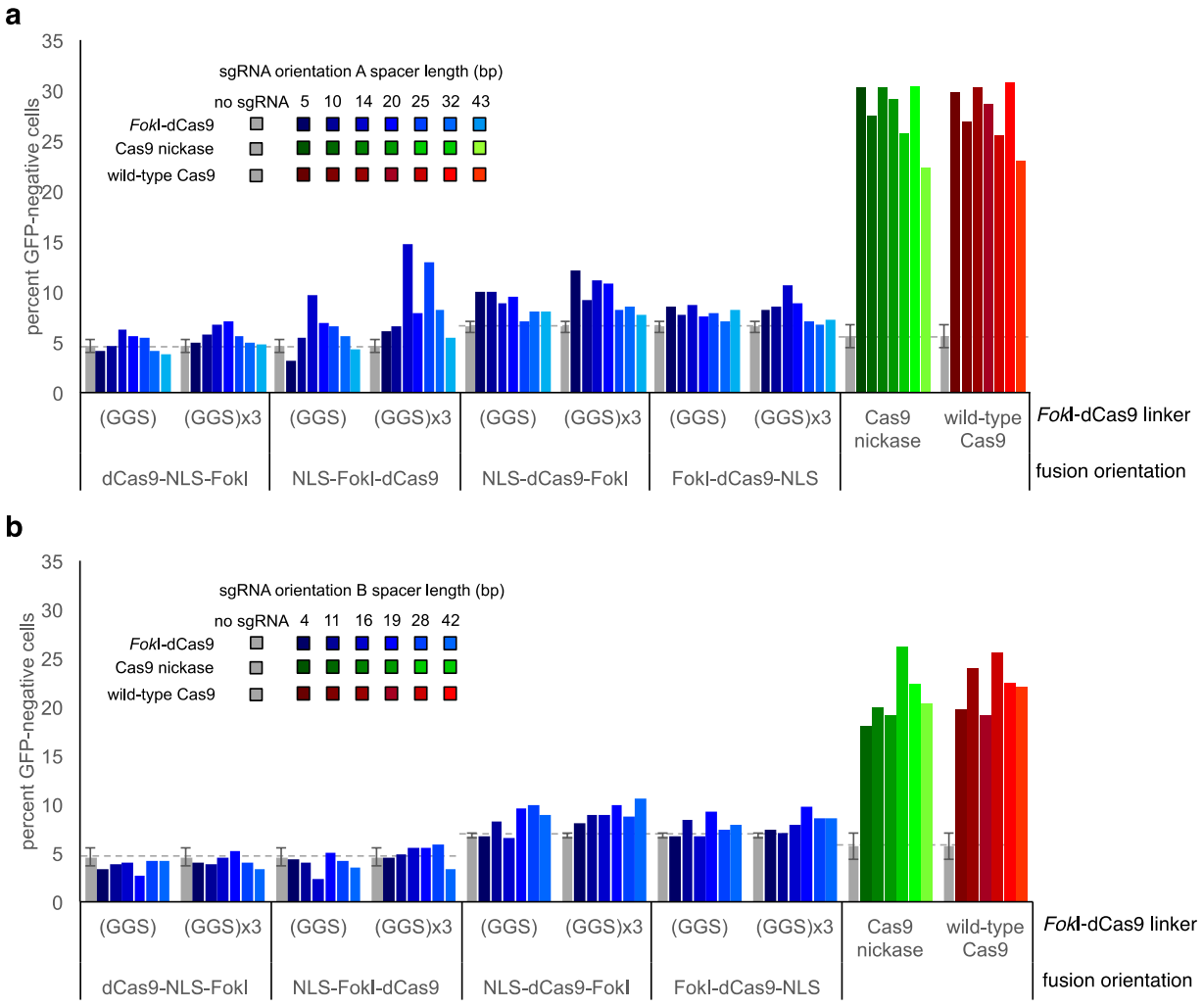

Figure 4.6 Activities of *Fok*I-dCas9 fusion candidates on sgRNA pairs of different orientations and of varying spacer lengths. Activity on (**a**) orientation A sgRNA spacers and (**b**) orientation B sgRNA spacers. All *Fok*I-dCas9 fusion data shown are the results of single trials. Wild-type Cas9 and Cas9 nickase data are the average of two replicates, while the 'no treatment' negative control data is the average of 6 replicates, with error bars representing one standard deviation. The gray dotted line across the Y-axis corresponds to the average of the 'no treatment' controls performed on the same day.

Next we optimized the protein linkers between the NLS and *Fok*I domain, and between the *Fok*I domain and dCas9 in the NLS-*Fok*I-dCas9 architecture. We tested 17 linkers with a wide range of amino acid compositions, predicted flexibilities, and lengths varying from 9 to 21 residues (**Figure 4.7**).

name	NLS-**linker**-Fok1	Fok1-**linker**-dCas9
*Fok*I-(GGS)x3	GGS	GGSGGSGGS
*Fok*I-(GGS)x6	GGS	GGSGGSGGSGGSGGSGGS
*Fok*I-L0	GGS	-
*Fok*I-L1	GGS	MKIIEQLPSA
*Fok*I-L2	GGS	VRHKLKRVGS
*Fok*I-L3	GGS	VPFLLEPDNINGKTC
*Fok*I-L4	GGS	GHGTGSTGSGSS
*Fok*I-L5	GGS	MSRPDPA
*Fok*I-L6	GGS	GSAGSAAGSGEF
*Fok*I-L7	GGS	SGSETPGTSESA
*Fok*I-L8	GGS	SGSETPGTSESATPES
*Fok*I-L9	GGS	SGSETPGTSESATPEGGSGGS
NLS-(GGS)	GGS	GGSM
NLS-(GGS)x3	GGSGGSGGS	GGSM
NLS-L1	VPFLLEPDNINGKTC	GGSM
NLS-L2	GSAGSAAGSGEF	GGSM
NLS-L3	SIVAQLSRPDPA	GGSM
wild-type Cas9	N/A	N/A
Cas9 nickase	N/A	N/A

Figure 4.7 Table of all linker variants tested. The initial active construct NLS-*Fok*I-dCas9 with a $(GGS)_3$ linker between *Fok*I and dCas9 was tested across a range of alternate linkers. The final choice of linkers for fCas9 is highlighted in blue. Wild-type Cas9 and Cas9 nickase were included for comparison.

Between the *Fok*I domain and dCas9 we identified a flexible 18-residue linker, $(GGS)_6$, and a 16-residue "XTEN" linker (*Fok*I-L8 in **Figure 4.7**) based on a previously reported engineered protein with an open, extended conformation,[26] as supporting the highest levels of genomic GFP modification (**Figure 4.8**). The XTEN protein was originally designed to extend the serum half-life of translationally fused biologic drugs by increasing their hydrodynamic radius, acting as protein-based functional analog to chemical PEGylation. Since XTEN is chemically stable, non-cationic, non-hydrophobic, and predicted to adopt an extended,

unstructured conformation, we hypothesized that an XTEN-based linker could functional as a stable, inert linker sequence for fusion proteins. The sequence of the XTEN protein tag from E-XTEN was analyzed, and repeating motifs within the amino acid sequence were aligned. The sequence used in the *Fok*I-dCas9 fusion construct *Fok*I-L8 (**Figure 4.7**) was derived from the consensus sequence of a common E-XTEN motif, and a 16-residue sequence was chosen from within this motif to test as a *Fok*I-dCas9 linker.

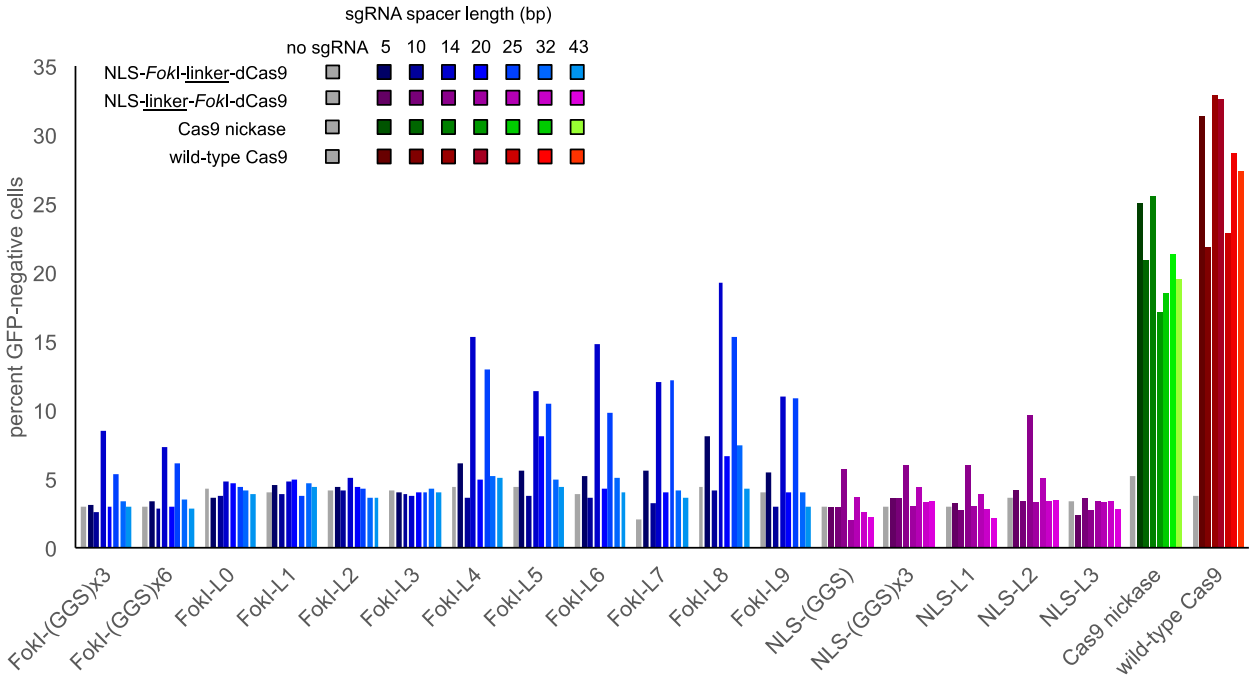

Figure 4.8 Optimization of protein linkers in NLS-*Fok*I-dCas9. The activity of *Fok*I-dCas9 fusions with linker variants. Each variant was tested across a range of spacer lengths from 5 to 43 bp using sgRNA pair orientation A. A control lacking sgRNA (grey) was included for each separate fusion construct.

Many of the *Fok*I-dCas9 linkers tested including the optimal XTEN linker resulted in nucleases with a marked preference for spacer lengths of ~15 and ~25 bp between half-sites, with all other spacer lengths, including 20 bp, showing substantially lower activity (**Figure 4.8**). This pattern of linker preference is consistent with a model in which the *Fok*I-dCas9 fusions must bind to opposite faces of the DNA double helix to cleave DNA, with optimal binding

taking place ~1.5 or 2.5 helical turns apart. The variation of the NLS-*Fok*I linkers did not strongly affect nuclease performance, especially when combined with the XTEN *Fok*I-dCas9 linker.

The NLS-GGS-*Fok*I-XTEN-dCas9 construct consistently exhibited the highest activity among the tested candidates, inducing loss of GFP in ~10% of cells over background, compared to ~15% and ~25% for Cas9 nickases and wild-type Cas9 nuclease, respectively (**Figure 4.9a**). All subsequent experiments were performed using this construct, hereafter referred to as fCas9. To confirm the ability of fCas9 to efficiently modify genomic target sites, we used the T7 endonuclease I Surveyor assay to measure the amount of mutation at each of seven target sites within the integrated GFP gene in HEK293 cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and either two distinct sgRNAs in orientation A or no sgRNAs as a negative control. Consistent with our flow cytometry-based studies, fCas9 was able to modify the GFP target sites with optimal spacer lengths of ~15 or ~25 bp at a rate of ~20%, comparable to the efficiency of nickase-induced modification and approximately two-thirds that of wild-type Cas9 (**Figure 4.9a-c)**.

Next we evaluated the ability of the optimized fCas9 to modify 14 distinct endogenous genomic loci in five genes by Surveyor assay. *AAVS1* (one site), *CLTA* (two sites), *EMX* (two sites), *HBB* (six sites) *VEGF* (three sites), and were targeted with two sgRNAs per site in orientation A spaced at various lengths (**Figure 4.10)**. Consistent with the results of the experiments targeting GFP, at appropriately spaced target half-sites fCas9 induced efficient modification of all five genes, ranging from 8% to 22% target chromosomal site modification (**Figure 4.11a-e**). Among the sgRNA spacer lengths resulting in the highest modification at each of the six genes targeted (including GFP), fCas9 induced on average14.9% (± 6.0% s.d.)

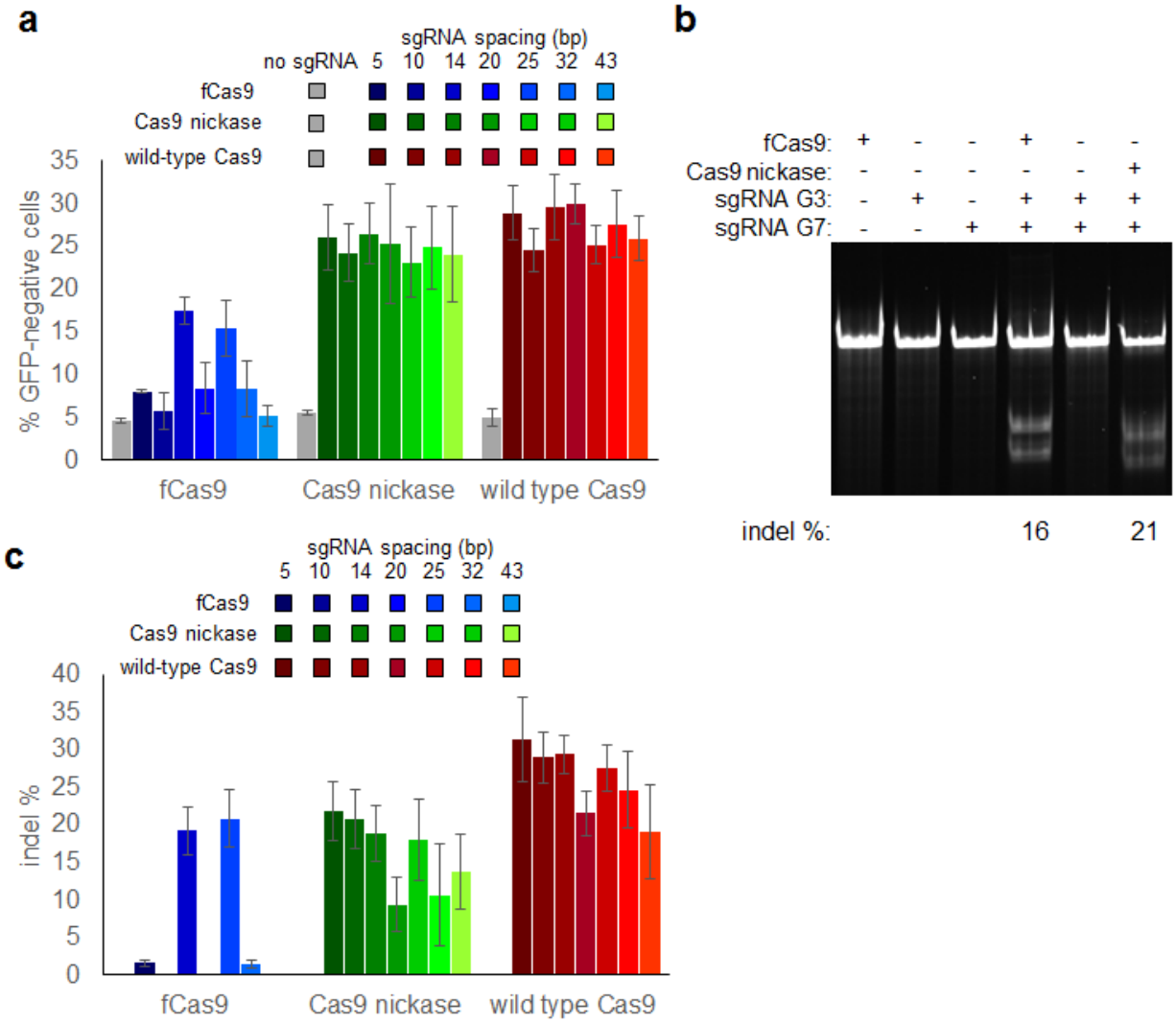

Figure 4.9 *GFP* gene modification by fCas9, Cas9 nickase, and wild-type Cas9. (**a**) GFP disruption activity measured by flow cytometry of cells treated transfected with fCas9, Cas9 nickase, or wild-type Cas9, with either no sgRNA, or sgRNA pairs of variable spacer length targeting the *GFP* gene in orientation A. (**b**) Indel modification efficiency from PAGE analysis of a Surveyor cleavage assay of renatured target-site DNA amplified from cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and two sgRNAs spaced 14 bp apart targeting the *GFP* site (sgRNAs G3 and G7; **Figure 4.4**), each sgRNA individually, or no sgRNAs. The indel modification percentage is shown below each lane for samples with modification above the detection limit (~2%). (**c**) Indel modification efficiency for treatments shown in (**a**).

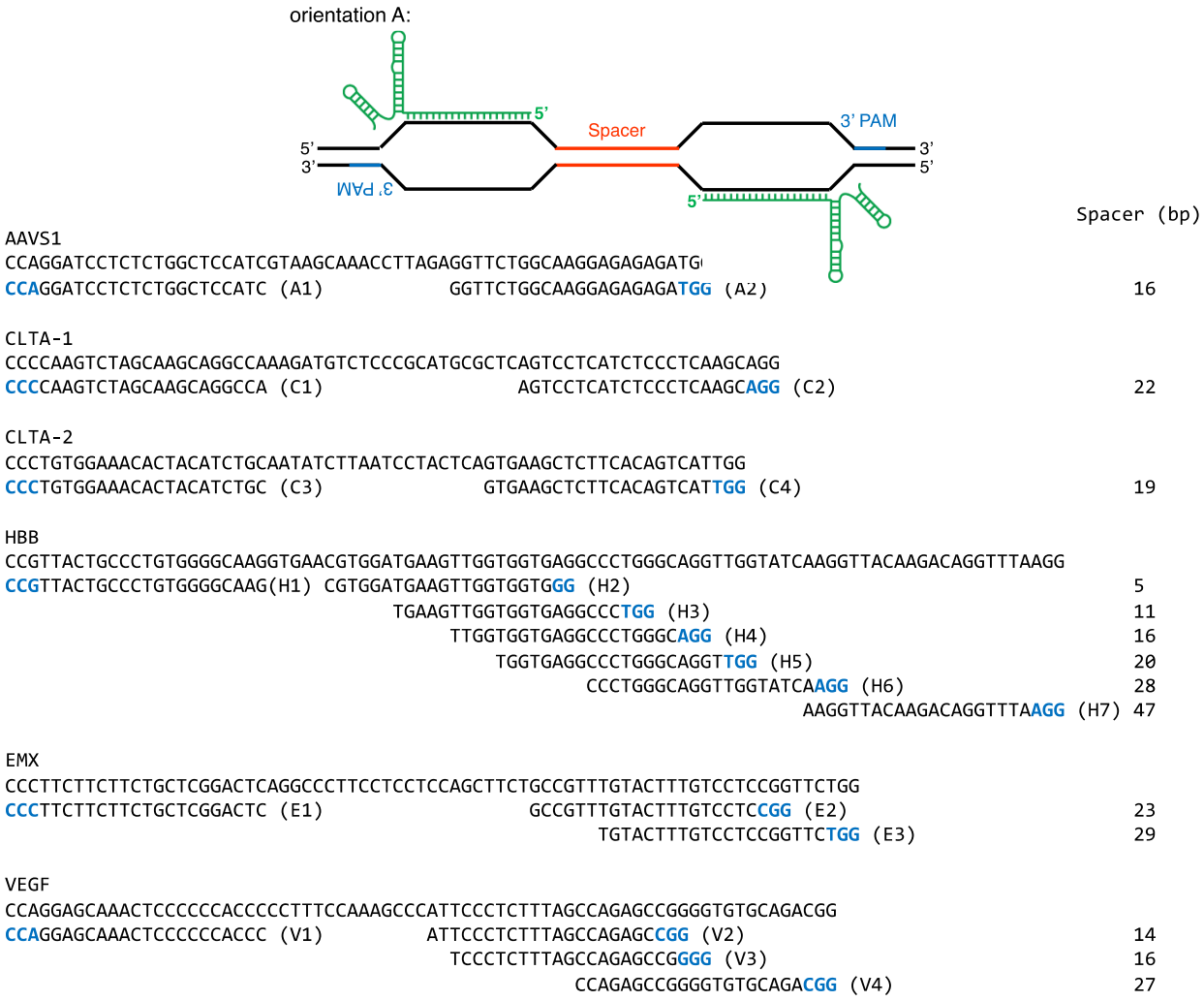

Figure 4.10 Target DNA sequences in endogenous human *AAVS1, CLTA, HBB, EMX*, and *VEGF* genes. sgRNA target sites tested within endogenous human *AAVS1*, *CLTA*, *HBB*, *EMX*, and *VEGF* genes. Fourteen paired sgRNA target sites were chosen to test the activity of the optimized fCas9 fusion in an orientation in which the PAM is distal from the cleaved spacer sequence (orientation A). Together, these 14 sgRNA pairs enabled testing of fCas9 fusion variants across twelve spacer lengths ranging from 5 to 47 bp.

modification, while Cas9 nickase and wild-type Cas9 induced on average 20.6% (± 5.6% s.d.) and 28.2% (± 6.2% s.d.) modification, respectively, from their optimal sgRNA pairs for each gene.

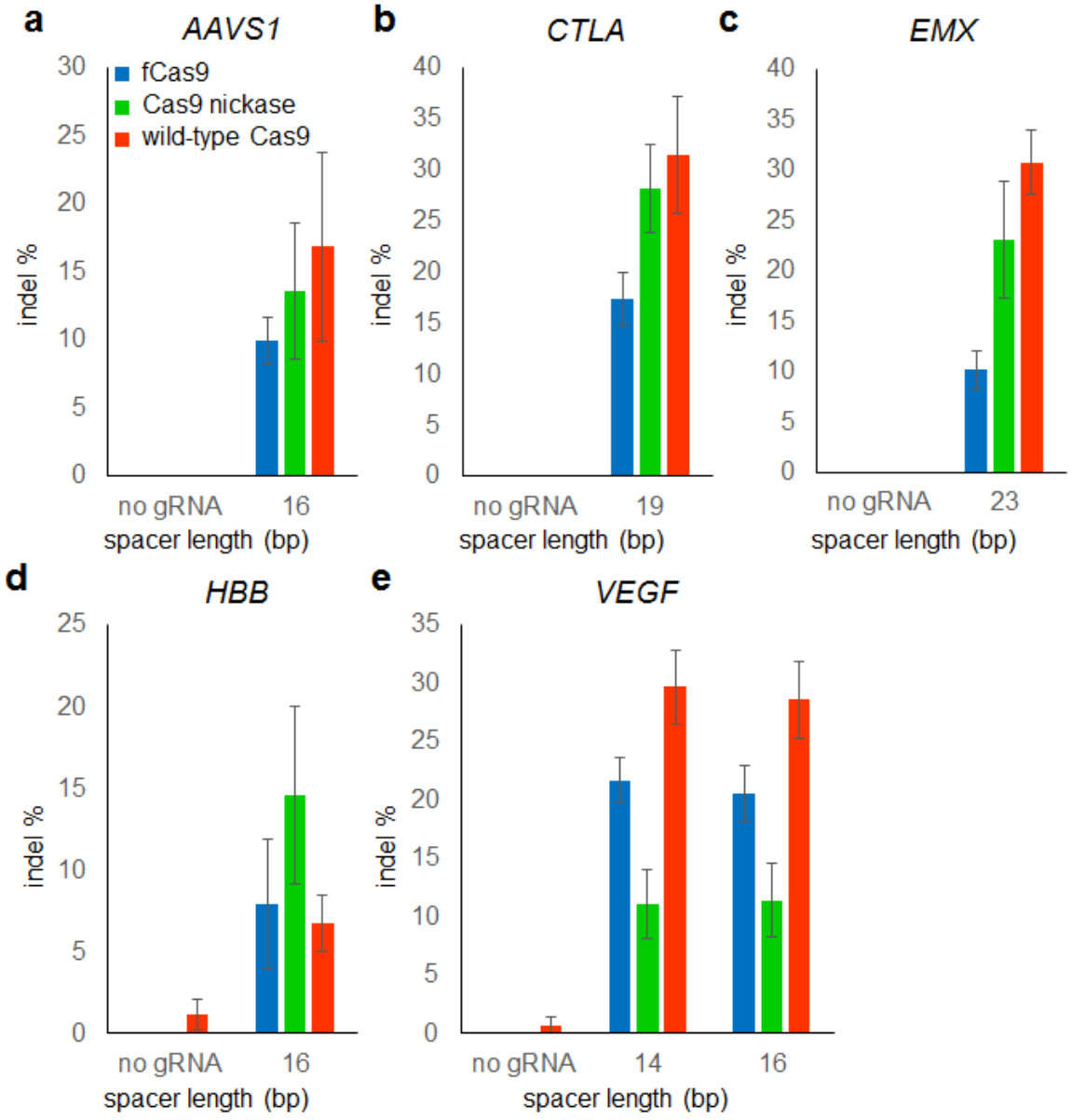

Figure 4.11 Genomic DNA modification by fCas9, Cas9 nickase, and wild-type Cas9. Indel modification efficiency as measured by Surveyor assay for target sites within the (**c**) *AAVS1* (**e**) *CLTA*, (**f**) *EMX,* (**g**) *HBB*, and (**h**) *VEGF* genes. Error bars reflect standard error of the mean from three biological replicates performed on different days.

As the sgRNA requirements of fCas9 potentially restrict the number of potential off-target substrates of fCas9, we compared the effect of guide RNA orientation on the ability of fCas9, Cas9 nickase, and wild-type Cas9 to cleave target GFP sequences. Consistent with

previous reports,[5–7] Cas9 nickase efficiently cleaved targets when guide RNAs were bound either in orientation A or orientation B, similar to wild-type Cas9 (**Figure 4.12a, b**). In contrast, fCas9 only cleaved the GFP target when guide RNAs were aligned in orientation A (**Figure 4.9a-c** and **Figure 4.13a, b**). This orientation requirement further limits opportunities for undesired off-target DNA cleavage.

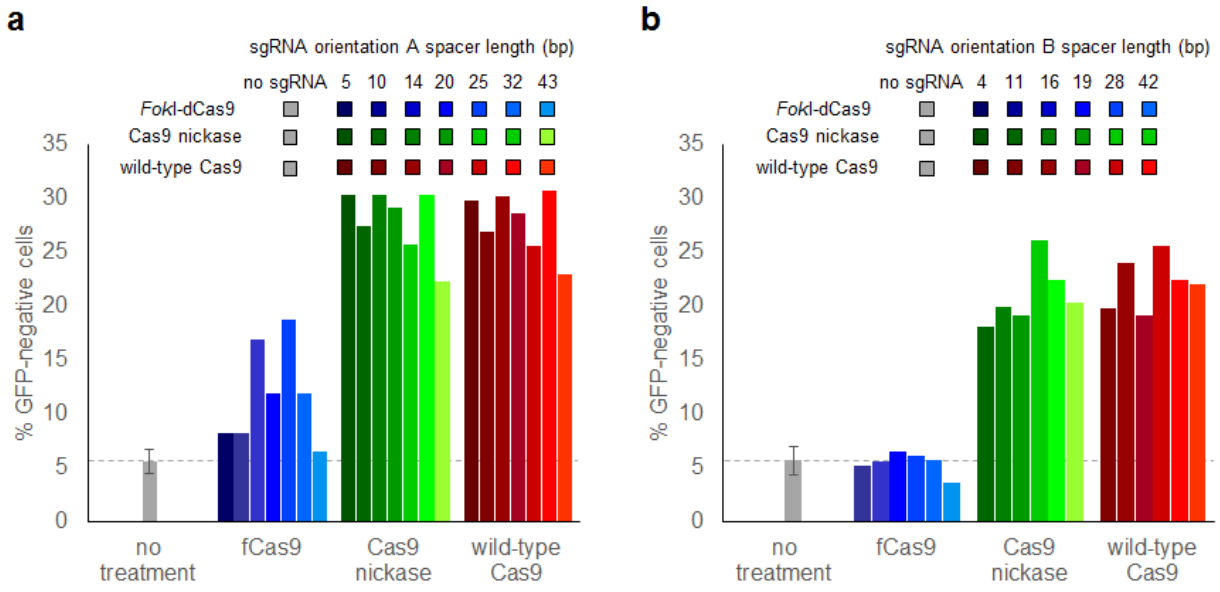

Figure 4.12 Dependence of fCas9 on sgRNA pair orientation. (**a**) GFP gene disruption by wild-type Cas9, Cas9 nickase, fCas9 using sgRNA pairs in orientation A. High activity of fCas9 requires spacer lengths of ~15 or 25 bp. (**b**) GFP gene disruption using sgRNA pairs in orientation B. Cas9 nickase, but not fCas9, accepts either orientation of sgRNA pairs. The "no treatment" control refers to cells receiving no plasmid DNA.

Importantly, no modification was observed by GFP disruption or Surveyor assay when any of four single sgRNAs were expressed individually with fCas9, as expected since two simultaneous binding events are required for *Fok*I activity (**Figure 4.9b** and **Figure 4.13**). In contrast, *GFP* gene disruption resulted from expression of any single sgRNA with wild-type Cas9 (as expected) and, in the case of two single sgRNAs, with Cas9 nickase. High-throughput

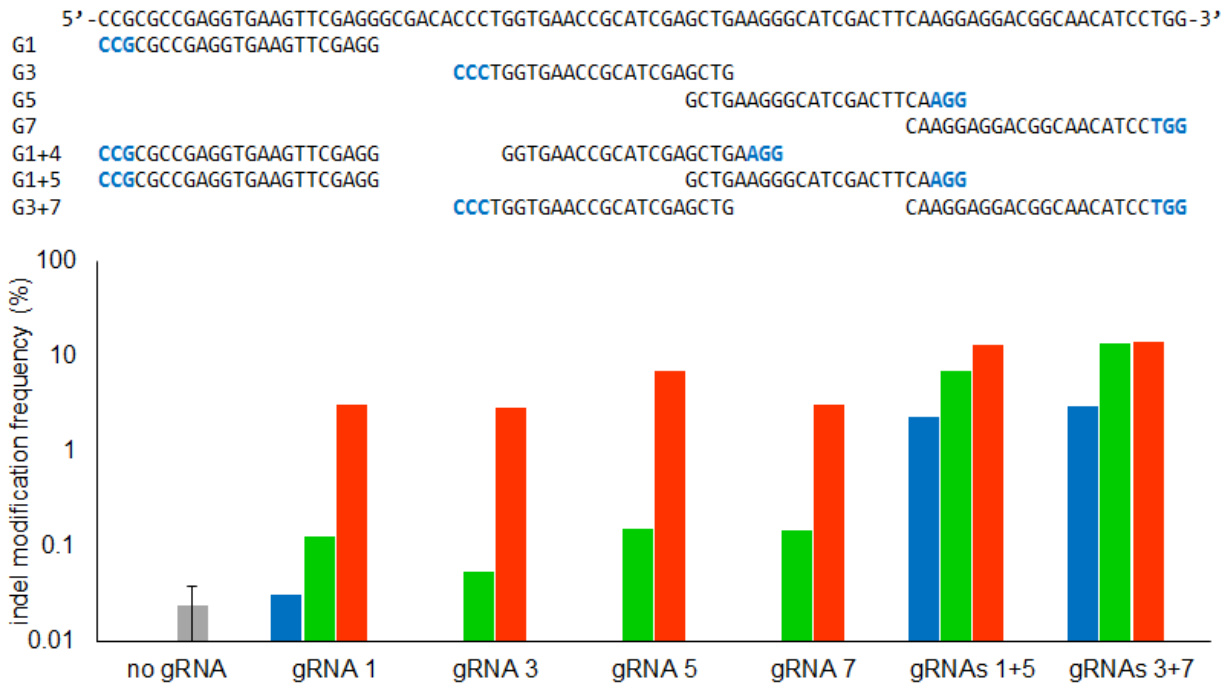

Figure 4.13. Paired sgRNA dependence of fCas9 compared to Cas9 nickase and wild-type Cas9. Indel frequency identified from high-throughput sequencing of *GFP* on-target sites amplified from genomic DNA isolated from human cells treated with a plasmid expressing wild-type Cas9, Cas9 nickase, or fCas9; and either a plasmid expressing a single sgRNA (G1, G3, G5, or G7), or two plasmids each expressing a different sgRNA (G1+G5 or G3+G7). As a negative control, transfection and sequencing were performed in triplicate as above without any sgRNA expression plasmids.

sequencing to detect indels at the *GFP* target site in cells treated with a single sgRNA and fCas9, Cas9 nickase, or wild-type Cas9 revealed the expected substantial level of modification ranging from 2.3% to 14.3% of sequence reads. Modification by fCas9 in the presence of any of the four single sgRNAs was not detected above background (< ~0.03% modification), consistent with the requirement of fCas9 to engage two sgRNAs in order to cleave DNA. In contrast, Cas9 nickases in the presence of single sgRNAs resulted in modification levels ranging from 0.05% to 0.16% at the target site (**Figure 4.13**). The detection of bona fide indels at target sites following Cas9 nickase treatment with single sgRNAs confirms the mutagenic potential of genomic DNA nicking, consistent with previous reports.[5,6,14,17,18] These results collectively demonstrate that

Cas9 nickase can induce genomic DNA modification in the presence of a single sgRNA, in contrast with the absence of single-sgRNA modification by fCas9. Taken together, these results indicate that fCas9 can modify genomic DNA efficiently and in a manner that requires simultaneous engagement of two guide RNAs targeting adjacent sites, unlike the ability of wild-type Cas9 and Cas9 nickase to cleave DNA when bound to a single guide RNA.

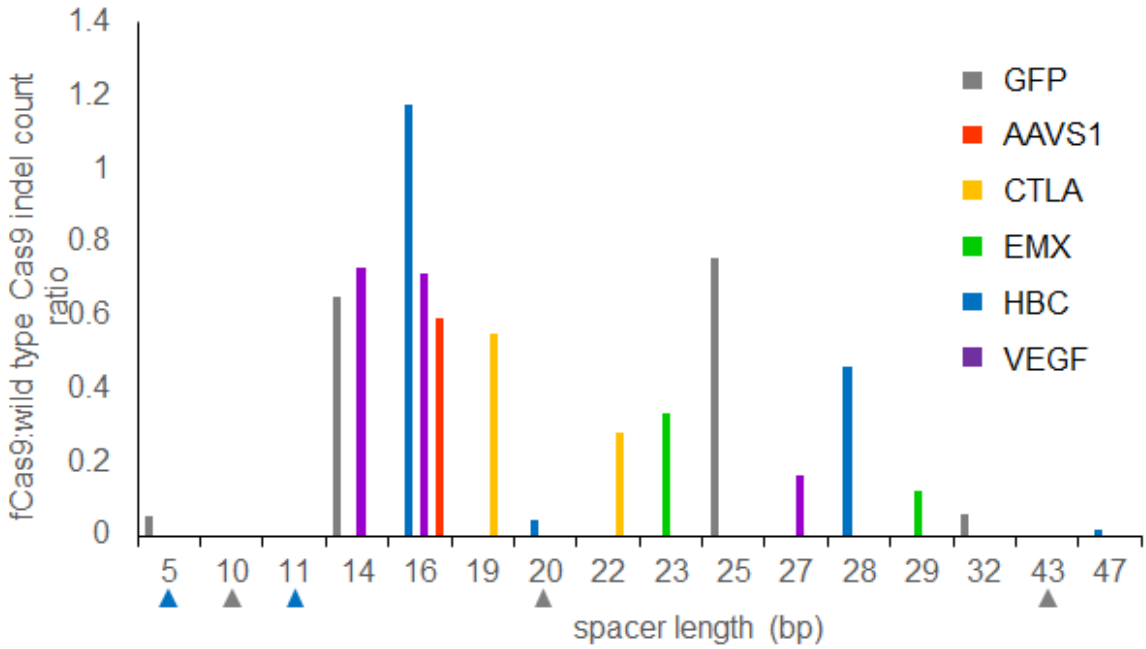

Figure 4.14. fCas9 indel frequency of genomic targets reflects sgRNA pair spacer length preference. The graph shows the relationship between spacer length (number of bp between two sgRNAs) and the indel modification efficiency of fCas9 normalized to the indel modification efficiency of the same sgRNAs co-expressed with wild-type Cas9 nuclease. Colored triangles below the X-axis denote spacer lengths that were tested but which yielded no detectable indels for the indicated target gene. These results suggest that fCas9 requires ~15 bp or ~25 bp between half-sites to efficiency cleave DNA.

The above results collectively reveal much more stringent spacer, sgRNA orientation, and guide RNA pairing requirements for fCas9 compared with Cas9 nickase. In contrast with fCas9 (**Figure 4.14)**, Cas9 nickase cleaved sites across all spacers assayed (5- to 47- bp in orientation A and 4 to 42 bp in orientation B in this work) (**Figures 4.9, 4.12,** and **4.13**). These

observations are consistent with previous reports of Cas9 nickases modifying sites targeted by sgRNAs with spacer lengths up to 100 bp apart.[6,7] The more stringent spacer and sgRNA orientation requirements of fCas9 compared with Cas9 nickase reduces the number of potential genomic off-target sites of the former by approximately 10-fold (**Appendix A)**. Although the more stringent spacer requirements of fCas9 also reduce the number of potential targetable sites, sequences that conform to the fCas9 spacer and dual PAM requirements exist in the human genome on average once every 34 bp (9.2×10^7 sites in 3.1×10^9 bp) (**Appendix A**). We also anticipate that the growing number of Cas9 homologs with different PAM specificities[28] will further increase the number of targetable sites using the fCas9 approach.

To evaluate the DNA cleavage specificity of fCas9, we measured the modification of known Cas9 off-target sites of *CLTA*, *EMX*, and *VEGF* genomic target sites.[1,2,7,8] The target site and its corresponding known off-target sites (**Appendix B**) were amplified from genomic DNA isolated from HEK293 cells treated with fCas9, Cas9 nickase, or wild-type Cas9 and two sgRNAs spaced 19 bp apart targeting the *CLTA* site, two sgRNAs spaced 23 bp apart targeting the *EMX* site, two sgRNAs spaced 14 bp apart targeting the *VEGF* site, or two sgRNAs targeting an unrelated site (GFP) as a negative control. In total 11 off-target sites were analyzed by high-throughput sequencing (**Figure 4.15a-d**). Sequences containing insertions or deletions of two or more base pairs in potential genomic off-target sites and present in significantly greater numbers (P value < 0.005, Fisher's exact test) in the target sgRNA-treated samples versus the control sgRNA-treated samples were considered Cas9 nuclease-induced genome modifications. For 10 of the 11 off-target sites assayed, fCas9 did not result in any detectable genomic off-target modification within the sensitivity limit of our assay (< 0.002%, see Methods), while

demonstrating substantial on-target modification efficiencies of 5% to 10% (**Figure 4.15a-c** and **Appendix C**)

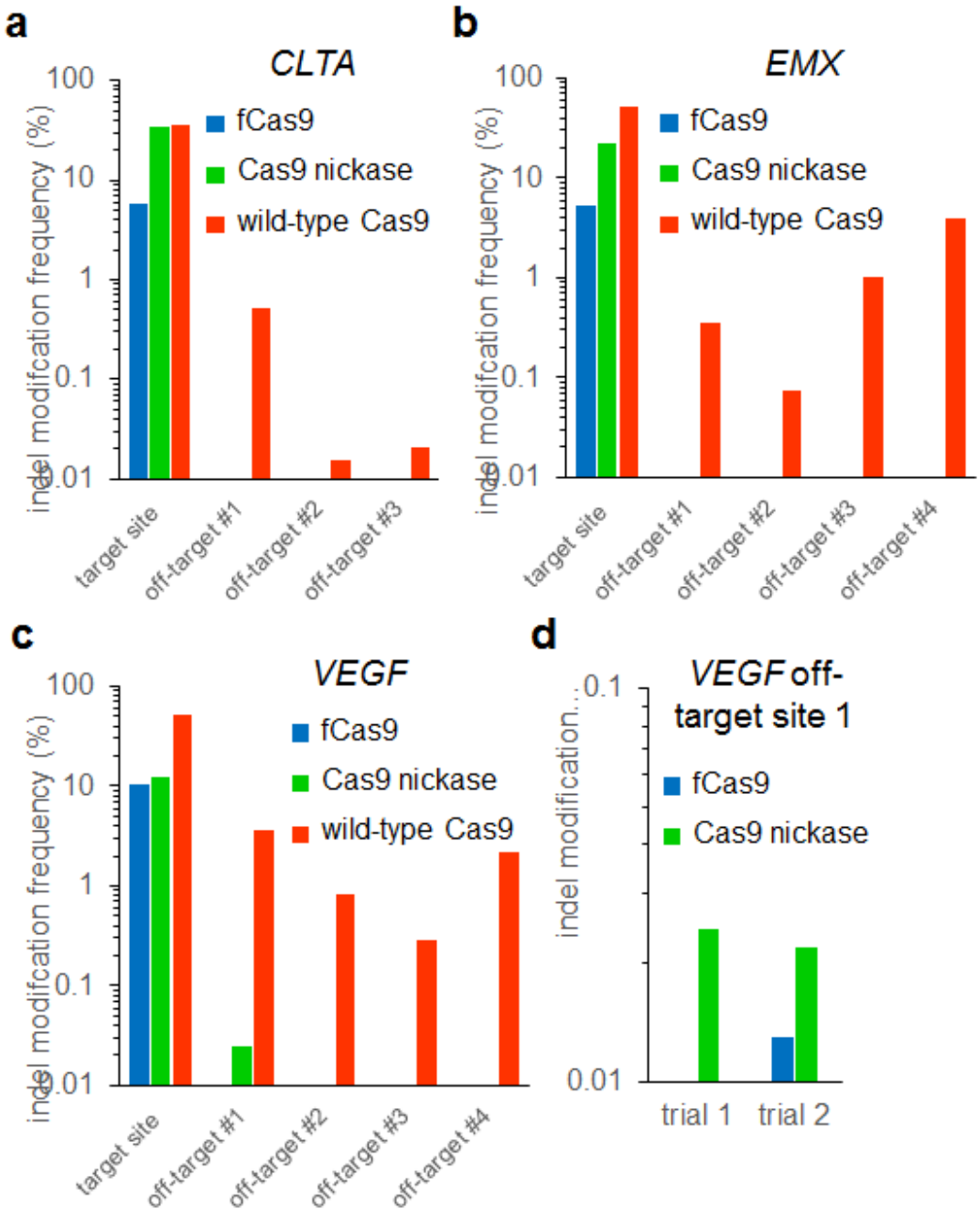

Figure 4.15 Off-target indel mutation frequency induced by fCas9, Cas9 nickase, or wild-type Cas9. Indels identified via high-throughput DNA sequencing of amplified genomic on-target sites and off-target sites from human cells treated with fCas9, Cas9 nickase, or wild-type Cas9. (**a**) *CLTA* site (sgRNAs C1 and C2), (**c**) *EMX* site (sgRNAs E1 and E2), or (**d, e**) *VEGF* site (sgRNAs V1 and V2). In (**a-d**), all significant (P value < 0.005 Fisher's Exact Test) indel frequencies are shown.

The detailed inspection of fCas9-modified *VEGF* on-target sequences (**Figure 4.16**) revealed a prevalence of deletions ranging from two to dozens of base pairs consistent with cleavage occurring in the DNA spacer between the two target binding sites. For each target site at *CLTA*, *EMX*, and *VEGF*, fCas9 predominantly induces deletions with insertions representing < 5% of all modifications. This prevalence of deletions is also observed with TALENs, which have similar spacer length preferences.[29]

```
fCas9 nuclease modifications of VEGF on-target site:
8959 CCAGGAGCAAACTCCCCCCACCCcctttccaaagcccATTCCCTCTTTAGCCAGAGCCGG (ref)
 125 ccaggagcaaactccccca------------agcccattccctctttagccagagccgg
 121 ccaggagcaaactccccccaccccctt-----------ttccctctttagccagagccgg
  77 ccaggagcaaactccccccaccccct------------ttccctctttagccagagccgg
  73 ccaggagcaaactccccca-------------gcccattccctctttagccagagccgg
  48 ccaggagcaaactccccccaccccc------------attccctctttagccagagccgg
  44 ccaggagcaaactccccccaccccttt--------------------agccagagccgg
  24 ccaggagcaaactccccccaccccttt--aaagcccattccctctttagccagagccgg
  22 ccaggagcaaactccccc-------------aagcccattccctctttagccagagccgg
```

Figure 4.16 Examples of modified sequences at the *VEGF* on-target site with fCas9. The unmodified genomic site is the first sequence, followed by the top eight sequences containing deletions. The numbers before each sequence indicate high-throughput sequencing counts. The sgRNA target sites are bold and capitalized.

In contrast, genomic off-target DNA cleavage was observed for wild-type Cas9 at all 11 sites assayed. Using the detection limit of the assay as an upper bound for off-target fCas9 activity, we calculated that fCas9 has a much lower off-target modification rate than wild-type Cas9 nuclease. At the 11 off-target sites modified by wild-type Cas9 nuclease, fCas9 resulted in on-target:off-target modification ratios at least 140-fold higher than that of wild-type Cas9 (**Figure 4.15a-d**).

Consistent with previous reports,[5,7,8] Cas9 nickase also induced substantially fewer off-target modification events (1/11 off-target sites modified at a detectable rate) compared to wild-type Cas9. An initial high-throughput sequencing assay revealed significant (P value < 10^{-3}, Fisher's Exact Test) modification induced by Cas9 nickases in 0.024% of sequences at *VEGF* off-target site 1. This genomic off-target site was not modified by fCas9 despite similar *VEGF* on-target modification efficiencies of 12.3% for Cas9 nickase and 10.4% for fCas9 (**Figure 4.15c** and **Appendix C**). Because Cas9 nickase-induced modification levels were within an order of magnitude of the limit of detection and fCas9 modification levels were undetected, we repeated the experiment with a larger input DNA samples and a greater number of sequence reads (150 versus 600 ng genomic DNA and > 8 x 10^5 versus > 23 x 10^5 reads for the initial and second trial, respectively) to detect off-target cleavage at this site by Cas9 nickase or fCas9. From this deeper interrogation, we observed Cas9 nickase and fCas9 to both significantly modify (P value < 10^{-5}, Fisher's Exact Test) *VEGF* off-target site 1 (**Figure 4.15d** and **Appendix C**). For both experiments interrogating the modification rates at *VEGF* off-target site 1, fCas9 exhibited a greater on-target:off-target DNA modification ratio than that of Cas9 nickase (> 5,150 and 1,650 for fCas9, versus 510 and 1,230 for Cas9 nickase, **Figure 4.15d**).

On either side of *VEGF* off-target site 1 there exist no other sites with six or fewer mutations from either of the two half-sites of the *VEGF* on-target sequence. We speculate that the first 11 bases of one sgRNA (V2) might hybridize to the single-stranded DNA freed by canonical Cas9:sgRNA binding within VEGF off-target site 1 (**Figure 4.10**). Through this sgRNA:DNA hybridization it is possible that a second Cas9 nickase or fCas9 could be recruited to modify this off-target site at a very rare, but detectable frequency. Judicious sgRNA pair design could eliminate this potential mode of off-target DNA cleavage, as *VEGF* off-target site 1

is highly unusual in its ability to form 11 consecutive potential base pairs with the second sgRNA of a pair. In general, fCas9 was unable to modify the genomic off-target sites tested because of the absence of any adjacent second binding site required to dimerize and induce cleavage by the *Fok*I nuclease domain.

Discussion

The optimized *Fok*I-dCas9 fusion architecture developed in this work modified all 11 genomic loci targeted with gRNA spaced ~15 bp or ~25 bp apart, demonstrating the generality of using fCas9 to induce genomic modification in human cells, although modification with fCas9 was less efficient than with wild-type Cas9. The use of fCas9 is straightforward, requiring only that PAM sequences be present with an appropriate spacing and orientation, and using the same sgRNAs as wild-type Cas9 or Cas9 nickases. The observed low off-target:on-target modification ratios of fCas9, > 140-fold lower than that of wild-type Cas9, likely arises from the distinct mode of action of dimeric *Fok*I, in which DNA cleavage proceeds only if two DNA sites are occupied simultaneously by two *Fok*I domains at a specified distance (here, ~15 bp or ~25 bp apart) and in a specific half-site orientation. The resulting unusually low off-target activity of fCas9 may enable applications of Cas9:sgRNA-based technologies that require a very high degree of target specificity, such as *ex vivo* or *in vivo* therapeutic modification of human cells. This work also provides a foundation for future studies to characterize in greater detail and further improve the DNA cleavage activity and specificity of fCas9 *in vitro* and *in vivo*. For example, the use of recently described orthogonal Cas9 homologs[28] coupled with obligate heterodimeric *Fok*I variants[30] may offer additional specificity gains.

Methods

Oligonucleotides and PCR

All oligonucleotides were purchased from Integrated DNA Technologies (IDT). Oligonucleotide sequences are listed in Supplementary Notes. PCR was performed with 0.4 µL of 2 U/µL Phusion Hot Start Flex DNA polymerase (NEB) in 50 µL with 1x HF Buffer, 0.2 mM dNTP mix (0.2 mM dATP, 0.2 mM dCTP, 0.2 mM dGTP, 0.2 mM dTTP) (NEB), 0.5 µM of each primer and a program of: 98 ˚C, 1 min; 35 cycles of [98 ˚C, 15 s; 65 ˚C, 15 s; 72 ˚C, 30 s] unless otherwise noted.

Construction of FokI-dCas9, Cas9 Nickase and sgRNA Expression Plasmids

The human codon-optimized *streptococcus pyogenes* Cas9 nuclease with NLS and 3xFLAG tag (Addgene plasmid 43861)[2] was used as the wild-type Cas9 expression plasmid. PCR products of wild-type Cas9 expression plasmid were assembled with Gibson Assembly Cloning Kit (New England Biolabs) to construct Cas9 and *Fok*I-dCas9 variants. Expression plasmids encoding a single sgRNA construct (sgRNA G1 through G13) were cloned as previously described.[2] Briefly, sgRNA oligonucleotides containing the 20-bp protospacer target sequence were annealed and the resulting 4-bp overhangs were ligated into BsmBI-digested sgRNA expression plasmid. For all cloning, 1 µl of ligation or assembly reaction was transformed into Mach1 chemically competent cells (Life Technologies). *Fok*I-dCas9 expression plasmids will be available from Addgene.

Modification of Genomic GFP

HEK293-GFP stable cells (GenTarget) were used as a cell line constitutively expressing an Emerald GFP gene (GFP) integrated on the genome. Cells were maintained in Dulbecco's modified Eagle medium (DMEM, Life Technologies) supplemented with 10% (vol/vol) fetal bovine serum (FBS, Life Technologies) and penicillin/streptomycin (1x, Amresco). 5 ×

10^4 HEK293-GFP cells were plated on 48-well collagen coated Biocoat plates (Becton Dickinson). One day following plating, cells at ~75% confluence were transfected with Lipofecatmine 2000 (Life Technologies) according to the manufacturer's protocol. Briefly, 1.5 μL of Lipofecatmine 2000 was used to transfect 950 ng of total plasmid (Cas9 expression plasmid plus sgRNA expression plasmids). 700 ng of Cas9 expression plasmid, 125 ng of one sgRNA expression plasmid and 125 ng of the paired sgRNA expression plasmid with the pairs of targeted sgRNAs listed in **Figure 4.4**. Separate wells were transfected with 1 μg of a near-infrared iRFP670 (Addgene plasmid 45457)[31] as a transfection control. 3.5 days following transfection, cells were trypsinized and resuspended in DMEM supplemented with 10% FBS and analyzed on a C6 flow cytometer (Accuri) with a 488 nm laser excitation and 520 nm filter with a 20 nm band pass. For each sample, transfections and flow cytometry measurements were performed once.

T7 Endonuclease I Surveyor Assays of Genomic Modifications

HEK293-GFP stable cells were transfected with Cas9 expression and sgRNA expression plasmids as described above. A single plasmid encoding two separate sgRNAs was transfected. For experiments titrating the total amount of expression plasmids (Cas9 expression + sgRNA expression plasmid), 700/250, 350/125, 175/62.5, 88/31 ng of Cas9 expression plasmid/ng of sgRNA expression plasmid were combined with inert carrier plasmid, pUC19 (NEB), as necessary to reach a total of 950 ng transfected plasmid DNA.

Genomic DNA was isolated from cells 2 days after transfection using a genomic DNA isolation kit, DNAdvance Kit (Agencourt). Briefly, cells in a 48-well plate were incubated with 40 μL of tryspin for 5 min at 37 °C. 160 uL of DNAdvance lysis solution was added and the solution incubated for 2 hr at 55 °C and the subsequent steps in the Agencourt DNAdvance kit protocol

were followed. 40 ng of isolated genomic DNA was used as template to PCR amplify the targeted genomic loci with flanking Surveyor primer pairs. PCR products were purified with a QIAquick PCR Purification Kit (Qiagen) and quantified with Quant-iT™ PicoGreen ® dsDNA Kit (Life Technologies). 250ng of purified PCR DNA was combined with 2 μL of NEBuffer 2 (NEB) in a total volume of 19 μL and denatured then re-annealed with thermocycling at 95 ˚C for 5 min, 95 to 85 ˚C at 2 ˚C/s; 85 to 20 ˚C at 0.2 ˚C/s. The re-annealed DNA was incubated with 1 μl of T7 Endonuclease I (10 U/μl, NEB) at 37 °C for 15 min. 10 μL of 50% glycerol was added to the T7 Endonuclease reaction and 12 μL was analyzed on a 5% TBE 18-well Criterion PAGE gel (Bio-Rad) electrophoresed for 30 min at 150 V, then stained with 1x SYBR Gold (Life Technologies) for 30 min. Cas9-induced cleavage bands and the uncleaved band were visualized on an AlphaImager HP (Alpha Innotech) and quantified using ImageJ software.[32] The peak intensities of the cleaved bands were divided by the total intensity of all bands (uncleaved + cleaved bands) to determine the fraction cleaved which was used to estimate gene modification levels as previously described.[33] For each sample, transfections and subsequent modification measurements were performed in triplicate on different days.

High-throughput Sequencing of Genomic Modifications

HEK293-GFP stable cells were transfected with Cas9 expression and sgRNA expression plasmids, 700 ng of Cas9 expression plasmid plus 250 ng of a single plasmid expression a pair of sgRNAs were transfected (high levels) and for just Cas9 nuclease, 88 ng of Cas9 expression plasmid plus 31 ng of a single plasmid expression a pair of sgRNAs were transfected (low levels). Genomic DNA was isolated as above and pooled from three biological replicates. 150 ng or 600 ng of pooled genomic DNA was used as template to amplify by PCR the on-target and off-target genomic sites with flanking HTS primer pairs. Relative amounts of crude PCR

products were quantified by gel electrophoresis and samples treated with different sgRNA pairs or Cas9 nuclease types were separately pooled in equimolar concentrations before purification with the QIAquick PCR Purification Kit (Qiagen). ~500 ng of pooled DNA was run a 5% TBE 18-well Criterion PAGE gel (BioRad) for 30 min at 200 V and DNAs of length ~125 bp to ~300 bp were isolated and purified by QIAquick PCR Purification Kit (Qiagen). Purified DNA was PCR amplified with primers containing sequencing adaptors, purified and sequenced on a MiSeq high-throughput DNA sequencer (Illumina) as described previously.[1]

Sensitivity limit of off-target cleavage assays

The sensitivity of the high-throughput sequencing method for detecting genomic off-target cleavage is limited by the amount genomic DNA (gDNA) input into the PCR amplification of each genomic target site. A 1 ng sample of human gDNA represents only ~330 unique genomes, and thus only ~330 unique copies of each genomic site are present. PCR amplification for each genomic target was performed on a total of 150 ng, 300 ng, or 600 ng of input gDNA, which provides amplicons derived from at most 50,000, 100,000 or 200,000 unique gDNA copies, respectively. Therefore, the high-throughput sequencing assay cannot detect rare genome modification events that occur at a frequency of less than 1 in 50,000 (0.002%), less than 1 in 100,000 (0.001%), or less than 1 in 200,000 (0.0005%), respectively.

Data Analysis

Illumina sequencing reads were filtered and parsed with scripts written in Unix Bash as outlined in **Appendix D**. DNA sequences will be deposited in NCBI's Sequencing Reads Archive (SRA). Sample sizes for sequencing experiments were maximized (within practical experimental considerations) to ensure greatest power to detect effects. Statistical analyses for

Cas9-modified genomic sites in **Appendix C** were performed as previously described[34] with multiple comparison correction using the Bonferroni method.

References

1. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31,** 839–843 (2013).

2. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–826 (2013).

3. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–832 (2013).

4. Cradick, T. J., Fine, E. J., Antico, C. J. & Bao, G. CRISPR/Cas9 systems targeting -globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41,** 9584–9592 (2013).

5. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24,** 132–141 (2013).

6. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31,** 833–838 (2013).

7. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154,** 1380–1389 (2013).

8. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* (2014). doi:10.1038/nbt.2808

9. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (2013).

10. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (2012).

11. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci.* **109,** E2579–E2586 (2012).

12. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343,** 84–87 (2013).

13. Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26,** 808–816 (2008).

14. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339,** 823–826 (2013).

15. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2,** e00471–e00471 (2013).

16. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10,** 957–963 (2013).

17. Ramirez, C. L. *et al.* Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. *Nucleic Acids Res.* **40,** 5560–5568 (2012).

18. Wang, J. *et al.* Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Res.* **22,** 1316–1326 (2012).

19. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31,** 397–405 (2013).

20. Vanamee, É. S., Santagata, S. & Aggarwal, A. K. FokI requires two specific DNA sites for cleavage. *J. Mol. Biol.* **309,** 69–78 (2001).

21. Maeder, M. L. *et al.* CRISPR RNA–guided activation of endogenous human genes. *Nat. Methods* **10,** 977–979 (2013).

22. Pattanayak, V., Ramirez, C. L., Joung, J. K. & Liu, D. R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat. Methods* **8,** 765–770 (2011).

23. Guilinger, J. P. *et al.* Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat. Methods* (2014). doi:10.1038/nmeth.2845

24. Nishimasu, H. *et al.* Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* (2014). doi:10.1016/j.cell.2014.02.001

25. Jinek, M. *et al.* Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science* (2014). doi:10.1126/science.1247997

26. Schellenberger, V. *et al.* A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat. Biotechnol.* **27,** 1186–1190 (2009).

27. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).

28. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10,** 1116–1121 (2013).

29. Kim, Y., Kweon, J. & Kim, J.-S. TALENs and ZFNs are associated with different mutation signatures. *Nat. Methods* **10,** 185–185 (2013).

30. Doyon, Y. *et al.* Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat Methods* **8,** 74–9 (2011).

31. Shcherbakova, D. M. & Verkhusha, V. V. Near-infrared fluorescent proteins for multicolor in vivo imaging. *Nat. Methods* **10,** 751–754 (2013).

32. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9,** 671–675 (2012).

33. Guschin, D. Y. *et al.* in *Eng. Zinc Finger Proteins* (Mackay, J. P. & Segal, D. J.) **649,** 247–256 (Humana Press, 2010).

34. Sander, J. D. *et al.* In silico abstraction of zinc finger nuclease cleavage profiles reveals an expanded landscape of off-target sites. *Nucleic Acids Res.* **41,** e181–e181 (2013).

Appendices

Appendix A

a

Spacer length (b)	Number of paired sgRNA sites in orientation A	Number of paired sgRNA sites in orientation B
-8	6874293	NC
-7	6785996	NC
-6	6984064	NC
-5	7023260	NC
-4	6487302	NC
-3	6401348	NC
-2	6981383	NC
-1	7230098	NC
0	7055143	NC
1	6598582	NC
2	6877046	NC
3	6971447	NC
4	6505614	5542549
5	6098107	5663458
6	6254974	6819289
7	6680118	6061225
8	7687598	5702252
9	6755736	7306646
10	6544849	6387485
11	6918186	6172852
12	6241723	5799496
13	6233385	7092283
14	6298717	7882433
15	6181422	7472725
16	6266909	6294684
17	6647352	6825904
18	6103603	6973590
19	5896092	6349456
20	6000683	5835825
21	5858015	6056352
22	6116108	6531913
23	5991254	6941816
24	6114969	6572849
25	6135119	5671641

b

Cas9 variant	Preferred spacer lengths (bp)	Total sites
fCas9	13 to 19, or 22 to 29, in orientation A	92354891
Cas9 nickase	-8 to 100 in orientation A 4 to 42 in orientation B	953048977

Appendix A (continued)

Appendix A. Paired sgRNA target site abundances for fCas9 and Cas9 nickase in the human genome. (**a**) Column 2 shows the number of sites in the human genome with paired sgRNA binding sites in orientation A allowing for a spacer length from -8 bp to 25 bp (column 1) between the two sgRNA binding sites. sgRNA binding sites in orientation A have the NGG PAM sequences distal from the spacer sequence ($CCNN_{20}$-spacer-$N_{20}NGG$). Column 3 shows the number of sites in the human genome with paired sgRNA binding sites in orientation B allowing for a spacer length from 4 to 25 bp (column 1) between the two sgRNA binding sites. sgRNA binding sites in orientation B have the NGG PAM sequences adjacent to the spacer sequence ($N_{20}NGG$ spacer $CCNN_{20}$) . NC indicates the number of sites in the human genome was not calculated. Negative spacer lengths refer to target sgRNA binding sites that overlap by the indicated number of base pairs. (**b**) Sum of the number of paired sgRNA binding sites in orientation A with spacer lengths of 13 to 19 bp, or 22 to 29 bp, the spacer preference of fCas9 (**Figure 4.12**). Sum of the number of paired sgRNA binding sites with spacer lengths of -8 bp to 100 bp in orientation A, or 4 to 42 bp in orientation B, the spacer preference of Cas9 nickases (4 to 42 bp in orientation B is based on **Figures 4.9** and **4.14**, and -8 bp to 100 bp in orientation A is based on previous reports).

Appendix B

	Genomic target site
EMX_On	GAGTCCGAGCAGAAGAAGAAGGG
EMX_Off1	GAGgCCGAGCAGAAGAAagACGG
EMX_Off2	GAGTCCtAGCAGgAGAAGAAGaG
EMX_Off3	GAGTCtaAGCAGAAGAAGAAGaG
EMX_Off4	GAGTtaGAGCAGAAGAAGAAAGG
VEG_On	GGGTGGGGGGAGTTTGCTCCTGG
VEG_Off1	GGaTGGaGGGAGTTTGCTCCTGG
VEG_Off2	GGGaGGGtGGAGTTTGCTCCTGG
VEG_Off3	cGGgGGaGGGAGTTTGCTCCTGG
VEG_Off4	GGGgaGGGGaAGTTTGCTCCTGG
CLT2_On	GCAGATGTAGTGTTTCCACAGGG
CLT2_Off1	aCAaATGTAGTaTTTCCACAGGG
CLT2_Off2	cCAGATGTAGTaTTcCCACAGGG
CLT2_Off3	ctAGATGaAGTGcTTCCACATGG

Appendix B. Known off-target substrates of Cas9 target sites in *EMX*, *VEGF*, and *CLTA*. List of genomic on-target and off-targets sites of the EMX, VEGF, and CLTA are shown with mutations from on-target in lower case and red. PAMs are shown in blue.

Appendix C

a

Nuclease type:	wt Cas9	wt Cas9	Cas9 nickase	fCas9	wt Cas9	Cas9 nickase	fCas9
sgRNA pair target:	*CLTA*	*CLTA*	*CLTA*	*CLTA*	*GFP*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	125	1000	1000	1000	1000	1000
CLTA Sites							
CLT2_On							
Indels	3528	1423	3400	575	3	13	5
Total	10000	10000	10000	10000	10000	10000	10000
Modified (%)	35.280	14.230	34.000	5.750	0.030	0.130	0.050
P-value	<1.0E-300	<1.0E-300	<1.0E-300	1.4E-163			
On:off specificity	1	1		1			
CLT2_Off1							
Indels	316	44	2	2	1	3	3
Total	60620	64755	71537	63079	93883	91306	82055
Modified (%)	0.521	0.068	0.003	0.003	<0.002	0.003	0.004
P-value	1.3E-126	2.1E-16					
On:off specificity	68	209		>2850			
CLT2_Off2							
Indels	11	5	3	1	1	1	2
Total	72596	51093	59632	35541	69114	64412	39978
Modified (%)	0.015	0.010	0.005	0.003	<0.002	<0.002	0.005
P-value	6.5E-03						
On:off specificity	2328	1454		>2850			
CLT2_Off3							
Indels	11	10	0	0	1	1	1
Total	52382	44212	54072	48668	55670	58707	54341
Modified (%)	0.021	0.023	<0.002	<0.002	<0.002	<0.002	<0.002
P-value	2.7E-03	3.5E-03					
On:off specificity	1680	629		>2850			

Appendix C (continued)

B

Nuclease type:	wt Cas9	wt Cas9	Cas9 nickase	fCas9	wt Cas9	Cas9 nickase	fCas9
sgRNA pair:	*EMX*	*EMX*	*EMX*	*EMX*	*GFP*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	125	1000	1000	1000	1000	1000
EMX Site							
EMX_On							
Indels	5111	2683	2267	522	0	0	2
Total	10000	10000	10000	10000	10000	10000	10000
Modified (%)	51.110	26.830	22.670	5.220	<0.002	<0.002	0.020
P-value	<1.0E-300	<1.0E-300	<1.0E-300	1.0E-154			
On:off specificity	1	1	1	1			
EMX_Off1							
Indels	386	122	7	1	4	9	7
Total	109787	83420	124564	88424	102817	90020	96526
Modified (%)	0.352	0.146	0.006	<0.002	0.004	0.010	0.007
P-value	1.3E-103	2.8E-37					
On:off specificity	145	183	>11222	>2584			
EMX_Off2							
Indels	74	58	3	6	3	0	4
Total	98568	94108	105747	78871	81717	79469	79193
Modified (%)	0.075	0.062	0.003	0.008	0.004	<0.002	0.005
P-value	3.2E-16	1.4E-12					
On:off specificity	681	435	>11222	>2584			
EMX_Off3							
Indels	736	178	20	14	12	11	17
Total	72888	65139	82348	59593	74341	73408	75080
Modified (%)	1.010	0.273	0.024	0.023	0.016	0.015	0.023
P-value	2.5E-202	3.1E-44					
On:off specificity	51	98	>11222	>2584			
EMX_Off4							
Indels	4149	620	3	3	6	7	5
Total	107537	91695	91368	91605	111736	119643	128088
Modified (%)	3.858	0.676	0.003	0.003	0.005	0.006	0.004
P-value	<1.0E-300	1.9E-202					
On:off specificity	13	40	>11222	>2584			

Appendix C (continued)

c

Nuclease type:	wt Cas9	wt Cas9	Cas9 nickase	fCas9	wt Cas9	Cas9 nickase	fCas9
sgRNA pair:	*VEGF*	*VEGF*	*VEGF*	*VEGF*	*GFP*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	125	1000	1000	1000	1000	1000
VEGF Sites							
VEG_On							
Indels	5253	2454	1230	1041	8	0	1
Total	10000	10000	10000	10000	10000	10000	10000
Modified (%)	52.530	24.540	12.300	10.410	0.080	<0.002	0.010
P-value	<1.0E-300	<1.0E-300	<1.0E-300	6.6E-286			
On:off specificity	1	1	1	1			
VEG_Off1							
Indels	2950	603	22	0	0	4	1
Total	82198	71163	90434	77557	74765	79738	74109
Modified (%)	3.589	0.847	0.024	<0.002	<0.002	0.005	<0.002
P-value	<1.0E-300	3.2E-188	2.5E-06				
On:off specificity	15	29	506	>5150			
VEG_Off2							
Indels	863	72	3	3	0	2	1
Total	102501	49836	119702	65107	54247	65753	61556
Modified (%)	0.842	0.144	0.003	0.005	<0.002	0.003	<0.002
P-value	3.5E-159	9.6E-24					
On:off specificity	62	170	>6090	>5150			
VEG_Off3							
Indels	260	33	3	2	3	1	0
Total	91277	83124	90063	84385	62126	68165	69811
Modified (%)	0.285	0.040	0.003	0.002	0.005	<0.002	<0.002
P-value	6.8E-54	1.0E-05					
On:off specificity	184	618	>6090	>5150			
VEG_Off4							
Indels	1305	149	3	2	3	2	4
Total	59827	41203	65964	57828	60906	61219	62162
Modified (%)	2.181	0.362	0.005	0.003	0.005	0.003	0.006
P-value	<1.0E-300	2.7E-54					
On:off specificity	24	68	>6090	>5150			

Appendix C (continued)

d

Nuclease type:	Cas9 nickase	fCas9	Cas9 nickase	fCas9
sgRNA pair:	*VEGF*	*VEGF*	*GFP*	*GFP*
Total expression plasmids (ng):	1000	1000	1000	1000
VEGF Sites				
VEG_On				
Indels	2717	2122	10	13
Total	10000	10000	10000	10000
Modified (%)	27.170	21.220	0.100	0.130
P-value	<1.0E-300	<1.0E-300		
On:off specificity	1	1		
VEG_Off1				
Indels	67	30	3	2
Total	302573	233567	204454	190240
Modified (%)	0.022	0.013		
P-value	5.9E-12	2.5E-06		
On:off specificity	1227	1652		

Appendix C. Cellular modification induced by wild-type Cas9, Cas9 nickase, and fCas9 at on-target and off-target genomic sites. (**a**) Results from sequencing *CLTA* on-target and previously reported genomic off-target sites amplified from 150 ng genomic DNA isolated from human cells treated with a plasmid expressing either wild-type Cas9, Cas9 nickase, or fCas9 and a single plasmid expressing two sgRNAs targeting the *CLTA* on-target site (sgRNA C3 and sgRNA C4). As a negative control, transfection and sequencing were performed as above, but using two sgRNAs targeting the *GFP* gene on-target site (sgRNA G1, G2 or G3 and sgRNA G4, G5, G6 or G7. Indels: the number of observed sequences containing insertions or deletions consistent with any of the three Cas9 nuclease-induced cleavage. Total: total number of sequence counts while only the first 10,000 sequences were analyzed for the on-target site sequences. Modified: number of indels divided by total number of sequences as percentages. Upper limits of potential modification were calculated for sites with no observed indels by assuming there is less than one indel then dividing by the total sequence count to arrive at an upper limit modification percentage, or taking the theoretical limit of detection (1/49,500), whichever value was larger. P-values: For wild-type Cas9 nuclease, Cas9 nickase or fCas9 nuclease, P-values were calculated as previously reported[18] using a two-sided Fisher's exact test between each sample treated with two sgRNAs targeting the *CLTA* on-target site and the control sample treated with two sgRNAs targeting the *GFP* on-target site. P-values of < 0.0045 were considered significant and shown based on conservative multiple comparison correction using the Bonferroni method. On:off specificity is the ratio of on-target to off-target genomic modification frequency for each site. (**b**) Experimental and analytic methods as in (**a**) applied to *EMX* target sites using a single plasmid expressing two sgRNAs targeting the *EMX* on-

target site (sgRNA E1 and sgRNA E2). (**c**) Experimental and analytic methods as in (**a**) applied to *VEGF* target sites using a single plasmid expressing two sgRNAs targeting the *VEGF* on-target site (sgRNA V1 and sgRNA v2). (d) Experimental and analytic methods as in (**a**) applied to *VEGF* on-target and *VEGF* off-target site 1 amplified from 600 ng genomic DNA to increase detection sensitivity to 1/198,000.

Appendix D

Computational Search for Potential Target Sites

1) The Patmatch program[4] was used to search the human genome (GRCh37/hg19 build) for pattern sequences corresponding to Cas9 binding sites (CCN N^{20} spacer N^{20}NGG for Orientation A and N^{20}NGG spacer CCN N^{20} for Orientation B)

Identification of Indels in Sequences of Genomic Sites

1) Sequence reads were initially filtered removing reads of less than 50 bases and removing reads with greater than 10% of the Illumina base scores not being B-J:

Example SeqA-1stread:

TTCTGAGGGCTGCTACCTGTACATCTGCACAAGATTGCCTTTACTCCATGCCTTTCTTCTTCTG
CTCTAACTCTGACAATCTGTCTTGCCATGCCATAAGCCCCTATTCTTTCTGTAACCCCAAGAT
GGTATAAAAGCATCAATGATTGGGC

Example SeqA-2stread:

AAAACTCAAAGAAATGCCCAATCATTGATGCTTTTATACCATCTTGGGGTTACAGAAAGAAT
AGGGGCTTATGGCATGGCAAGACAGATTGTCAGAGTTAGAGCAGAAGAAGAAAGGCATGG
AGTAAAGGCAATCTTGTGCAGATGTACAGGTAA

2) Find **the first 20 bases four bases from the start** of the reverse complement of SeqA-2ndread in SeqA-1stread allowing for 1 mismatch:

Reverse complement of SeqA-2ndread:

TTAC**CTGTACATCTGCACAAGATT**GCCTTTACTCCATGCCTTTCTTCTTCTGCTCTAACTCTG
ACAATCTGTCTTGCCATGCCATAAGCCCCTATTCTTTCTGTAACCCCAAGATGGTATAAAAGC
ATCAATGATTGGGCATTTCTTTGAGTTTT

Position in SeqA-1stread
TTCTGAGGGCTGCTAC**CTGTACATCTGCACAAGATTG**CCTTTACTCCATGCCTTTCTTCTTC
TGCTCTAACTCTGACAATCTGTCTTGCCATGCCATAAGCCCCTATTCTTTCTGTAACCCCAAG
ATGGTATAAAAGCATCAATGATTGGGC

3) Align and then combine sequences, removing any sequence with greater than 5% mismatches in the simple base pair alignment:

Combination of SeqA-1stread and SeqA-2ndread:

Appendix D (continued)

TTCTGAGGGCTGCTAC**CTGTACATCTGCACAAGATT**GCCTTTACTCCATGCCTTTCTTCTTC
TGCTCTAACTCTGACAATCTGTCTTGCCATGCCATAAGCCCCTATTCTTTCTGTAACCCCAAG
ATGGTATAAAAGCATCAATGATTGGGCATTTCTTTGAGTTTT

4) To identify the target site the flanking genomic sequences were searched for with the Patmatch program[4] allowing for varying amounts of bases from 1 to 300 between the flanking genomic sequences:

Target Site	Downstream genomic sequence	Upstream genomic sequence
EMX_On	GGCCTGCTTCGTGGCAATGC	ACCTGGGCCAGGGAGGGAGG
EMX_Off1	CTCACTTAGACTTTCTCTCC	CTCGGAGTCTAGCTCCTGCA
EMX_Off2	TGGCCCCAGTCTCTCTTCTA	CAGCCTCTGAACAGCTCCCG
EMX_Off3	TGACTTGGCCTTTGTAGGAA	GAGGCTACTGAAACATAAGT
EMX_Off4	TGCTACCTGTACATCTGCAC	CATCAATGATTGGGCATTTC
VEG_On	ACTCCAGTCCCAAATATGTA	ACTAGGGGGCGCTCGGCCAC
VEG_Off1	CTGAGTCAACTGTAAGCATT	GGCCAGGTGCAGTGATTCAT
VEG_Off2	TCGTGTCATCTTGTTTGTGC	GGCAGAGCCCAGCGGACACT
VEG_Off3	CAAGGTGAGCCTGGGTCTGT	ATCACTGCCCAAGAAGTGCA
VEG_Off4	TTGTAGGATGTTTAGCAGCA	ACTTGCTCTCTTTAGAGAAC
CLT2_On	CTCAAGCAGGCCCCGCTGGT	TTTTGGACCAAACCTTTTTG
CLT2_Off1	TGAGGTTATTTGTCCATTGT	TAAGGGGAGTATTTACACCA
CLT2_Off2	TCAAGAGCAGAAAATGTGAC	CTTGCAGGGACCTTCTGATT
CLT2_Off3	TGTGTGTAGGACTAAACTCT	GATAGCAGTATGACCTTGGG
SiteA_On	GCTCTGCCACCACAAGCTTTGGGCA	CCCTTTGCATCCATTCCCCCTACCA
SiteA_Off1	GGAGATGAACCAGCCTGCAGTCAAG	ACTGATCTATGCCTGTGCCTTTGTG
SiteA_Off2	CCCAGTCCCTATCACAAAAAAAGAT	ACATTGATCATCATGGCCACTGGAT
SiteA_Off3	TCCTGATGCCAGCACTCAGTGCCTG	AAGAGCACCAAGTACAGTCTGTGGC
SiteB_On	TTCCCAAACTGCTGGGATTACAGGC	TGCTACTGTGTACTAAGGGCATAGT
SiteB_Off1	CTCAGCCTCTCAAAGTGCTGGGATT	TATCTCCTTCCCTTTCCTTCCCTTC
SiteB_Off2	CTCCCAAAGTGCTGGGATTACAGGC	TTTGGTTTATAGAAACACCATTGAT
SiteB_Off3	CTCCCAAAGTGCTGGGATTACAAGG	GAATGTTAAGTTTGTCCAGAGGCCA
SiteB_Off4	GCCTCCCAAAGTGCTGGGATTACAG	CCAGCACTTTGGGAGGCCAAAGCGG
SiteC_On	CCTCAGCTTCCCAAACTGCTGAGAT	GTGTGACCTTTGCTTTGGAACTGTG
SiteC_Off1	TCTCGACCTCCCTAAGTGCTGGGAT	CTTGCAGAAGAGTGCCAGTTGTGGT
SiteC_Off2	AATCTGCCCACCTCGGCCTCCCAAA	TACCACTTTTAAAATTTCACTTCTC
SiteC_Off3	CTCCCAAAGTGCTGGTATTACAGGT	CTTTTGTCTTAATAATTCCTCTATT
SiteC_Off4	CTGCCTCAGCCTCCCGAAGTGCTAG	ATAATCCCAGCACTTTGAAAGGCTG

4) Any target site sequences corresponding to the same size as the reference genomic site in the human genome (GRCh37/hg19 build) were considered unmodified and any sequences not the

reference size were considered potential insertions or deletions. Sequences not the reference size were aligned with

ClustalW to the reference genomic site. Aligned sequences with more than one insertion or one deletion in the DNA spacer sequence in or between the two half-site sequences were considered indels. Since high-throughput sequencing can result in insertions or deletions of one base pairs (mis-phasing) at a low but relevant rates - indels of two bp are more likely to arise from Cas9 induced modifications.